The Unscrambler

Methods

By CAMO Software AS

CAMO

www.camo.com

This manual was produced using ComponentOne Doc-To-Help® 2005 together with Microsoft® Word. Visio and Excel were used to make some of the illustrations. The screen captures were taken with Paint Shop Pro.

Trademark Acknowledgments

Doc-To-Help® is a trademark of ComponentOne LLC.

Microsoft® is a registered trademark and Windows® 95, Windows® 98, Windows® NT, Windows® 2000, Windows® ME, Windows® XP, Excel and Word are trademarks of Microsoft Corporation.

PaintShop Pro is a trademark of JASC, Inc.

Visio is a trademark of Shapeware Corporation.

Restrictions

Information in this manual is subject to change without notice. No part of the documents that build it up may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of CAMO Software AS.

Software Version

This manual is up to date for version 9.6 of The Unscrambler®.

Document last updated on June 5, 2006.

Copyright © 1996-2006 CAMO Software AS. All rights reserved.

Contents

What Is New in The Unscrambler 9.6? 1

If You Are Upgrading from Version 9.5 1
If You Are Upgrading from Version 9.2 2
If You Are Upgrading from Version 9.1 3
If You Are Upgrading from Version 8.0.5 4
If You Are Upgrading from Version 8.0 5
If You Are Upgrading from Version 7.8 5
If You Are Upgrading from Version 7.6 6
If You Are Upgrading from Version 7.5 7
If You Are Upgrading from Version 7.01 8

What is The Unscrambler? 11

Make Well-Designed Experimental Plans 11
Reformat, Transform and Plot your Data 12
Study Variations among One Group of Variables 12
Study Relations between Two Groups of Variables 13
Validate your Multivariate Models with Uncertainty Testing 13
Make Calibration Models for Three-way Data 13
Estimate New, Unknown Response Values 14
Classify Unknown Samples 14
Reveal Groups of Samples 14

Data Collection and Experimental Design 15

Principles of Data Collection and Experimental Design 15
- Data Collection Strategies 15
- What Is Experimental Design? 16
- Various Types of Variables in Experimental Design 16
- Investigation Stages and Design Objectives 18
- Designs for Unconstrained Screening Situations 19
- Designs for Unconstrained Optimization Situations 23
- Designs for Constrained Situations, General Principles 25
- Designs for Simple Mixture Situations 30
- Introduction to the D-Optimal Principle 35
- D-Optimal Designs Without Mixture Variables 37
- D-Optimal Designs With Mixture Variables 38
- Various Types of Samples in Experimental Design 39
- Sample Order in a Design 43
- Extending a Design 44
- Building an Efficient Experimental Strategy 47
- Advanced Topics for Unconstrained Situations 48
- Advanced Topics for Constrained Situations 49

Three-Way Data: Specific Considerations 52
- What Is A Three-Way Data Table? 52

Logical organization Of Three-Way Data Arrays 52
Unfolding Three-Way Data 53
Experimental Design and Data Entry in Practice 55
Various Ways To Create A Data Table 55
Build A Non-designed Data Table 56
Build An Experimental Design 57
Import Data 57
Save Your Data 57
Work With An Existing Data Table 57
Print Your Data 57

Represent Data with Graphs 59

The Smart Way To Display Numbers 59
Various Types of Plots 59
Line Plot 60
2D Scatter Plot 61
3D Scatter Plot 61
Matrix Plot 62
Normal Probability Plot 62
Histogram Plot 63
Plotting Raw Data 63
Line Plot of Raw Data 63
2D Scatter Plot of Raw Data 65
3D Scatter Plot of Raw Data 65
Matrix Plot of Raw Data 66
Normal Probability Plot of Raw Data 66
Histogram of Raw Data 67
Special Cases 69
Special Plots 69
Table Plot 69

Re-formatting and Pre-processing 71

Principles of Data Pre-processing 71
Filling Missing Values 72
Computation of Various Functions 72
Smoothing 72
Normalization 74
Spectroscopic Transformations 76
Multiplicative Scatter Correction 77
Adding Noise 78
Derivatives 78
Standard Normal Variate 81
Averaging 82
Transposition 82
Shifting Variables 82
User-Defined Transformations 82
Centering 82
Weighting 83
Pre-processing of Three-way Data 85
Re-formatting and Pre-processing in Practice 85
Make Simple Changes In The Editor 85
Organize Your Samples And Variables Into Sets 87
Change the Layout or Order of Your Data 87
Apply Transformations 87

Undo and Redo 88
Re-formatting and Pre-processing: Restrictions for 3D Data Tables 88
Re-formatting and Pre-processing: Restrictions for Mixture and D-Optimal Designs 89

Describe One Variable At A Time 91

Simple Methods for Univariate Data Analysis 91
Descriptive Statistics 91
First Data Check 91
Descriptive Variable Analysis 92
Plots For Descriptive Statistics 92
Univariate Data Analysis in Practice 92
Display Descriptive Statistics In The Editor 92
Study Your Variables Graphically 92
Compute And Plot Detailed Descriptive Statistics 93

Describe Many Variables Together 95

Principles of Descriptive Multivariate Analysis (PCA) 95
Purposes Of PCA 95
How PCA Works (In Short) 95
Calibration, Validation and Related Samples 97
Main Results Of PCA 97
More Details About The Theory Of PCA 99
How To Interpret PCA Results 100
PCA in Practice 102
Run A PCA 103
Save And Retrieve PCA Results 103
View PCA Results 103
Run New Analyses From The Viewer 104
Extract Data From The Viewer 105
How to Run an Analysis on 3-D Data 106

Combine Predictors and Responses In A Regression Model 107

Principles of Predictive Multivariate Analysis (Regression) 107
What Is Regression? 107
Multiple Linear Regression (MLR) 109
Principal Component Regression (PCR) 109
PLS Regression 110
Calibration, Validation and Related Samples 110
Main Results Of Regression 111
More Details About Regression Methods 114
How To Interpret Regression Results 115
Multivariate Regression in Practice 116
Run A Regression 116
Save And Retrieve Regression Results 117
View Regression Results 117
Run New Analyses From The Viewer 118
Extract Data From The Viewer 119

Validate A Model 121

Principles of Model Validation 121
What Is Validation? 121
Test Set Validation 121
Cross Validation 122

Leverage Correction 122
Validation Results 122
When To Use Which Validation Method 123
Uncertainty Testing With Cross Validation 123
How Does Martens' Uncertainty Test Work? 124
Application Example 125
More Details About The Uncertainty Test 129
Model Validation in Practice 130
How To Validate A Model 131
How To Display Validation Results 131
How To Display Uncertainty Test Results 132

Make Predictions 133

Principles of Prediction on New Samples 133
When Can You Use Prediction? 133
How Does Prediction Work? 133
Main Results Of Prediction 134
Prediction in Practice 135
Run A Prediction 135
Save And Retrieve Prediction Results 135
View Prediction Results 135

Classification 137

Principles of Sample Classification 137
SIMCA Classification 137
Main Results of Classification 138
Outcomes Of A Classification 140
Classification And Regression 140
Classification in Practice 141
Run A Classification 141
Save And Retrieve Classification Results 142
View Classification Results 142
Run A PLS Discriminant Analysis 143

Clustering 145

Principles of Clustering 145
Distance Types 145
Quality of the Clustering 146
Main Results of Clustering 147
Clustering in Practice 147
Run A Clustering 147
View Clustering Results 147

Analyze Results from Designed Experiments 149

Specific Methods for Analyzing Designed Data 149
Simple Data Checks and Graphical Analysis 149
Study Main Effects and Interactions 149
Make a Response Surface Model 152
Analyze Results from Constrained Experiments 154
Analyzing Designed Data in Practice 157
Run an Analysis on Designed Data 157
Save And Retrieve Your Results 157
Display Data Plots and Descriptive Statistics 158

View Analysis of Effects Results 158
View Response Surface Results 159
View Regression Results for Designed Data 160

Multivariate Curve Resolution 161

Principles of Multivariate Curve Resolution (MCR) 161
What is MCR? 161
Data Suitable for MCR 161
Purposes of MCR 162
Main Results of MCR 163
More Details About MCR 165
How To Interpret MCR Results 169
Multivariate Curve Resolution in Practice 172
Run An MCR 173
Save And Retrieve MCR Results 173
View MCR Results 173
Run New Analyses From The Viewer 174
Extract Data From The Viewer 175

Three-way Data Analysis 177

Principles of Three-way Data Analysis 177
From Matrices and Tables to Three-way Data 177
Notation of Three-way Data 178
Three-way Regression 181
Main Results of Tri-PLS Regression 183
Interpretation of a Tri-PLS Model 184
Three-way Data Analysis in Practice 184
Run A Tri-PLS Regression 185
Save And Retrieve Tri-PLS Regression Results 185
View Tri-PLS Regression Results 185
Run New Analyses From The Viewer 186
Extract Data From The Viewer 186
How to Run Other Analyses on 3-D Data 186

Interpretation Of Plots 187

Line Plots 187
Detailed Effects (Line Plot) 187
Discrimination Power (Line Plot) 187
Estimated Concentrations (Line Plot) 187
Estimated Spectra (Line Plot) 188
F-Ratios of the Detailed Effects (Line Plot) 188
Leverages (Line Plot) 188
Loadings for the X-variables (Line Plot) 189
Loadings for the Y-variables (Line Plot) 190
Loading Weights (Line Plot) 191
Mean (Line Plot) 191
Model Distance (Line Plot) 191
Modeling Power (Line Plot) 191
Predicted and Measured (Line Plot) 192
p-values of the Detailed Effects (Line Plot) 192
p-values of the Regression Coefficients (Line Plot) 192
Regression Coefficients (Line Plot) 192

Regression Coefficients with t-values (Line Plot) 193
RMSE (Line Plot) 194
Sample Residuals, MCR Fitting (Line Plot) 194
Sample Residuals, PCA Fitting (Line Plot) 194
Sample Residuals, X-variables (Line Plot) 194
Sample Residuals, Y-variables (Line Plot) 195
Scores (Line Plot) 195
Standard Deviation (Line Plot) 196
Standard Error of the Regression Coefficients (Line Plot) 196
Total Residuals, MCR Fitting (Line Plot) 196
Total Residuals, PCA Fitting (Line Plot) 197
Total Variance, X-variables (Line Plot) 197
Total Variance, Y-variables (Line Plot) 198
Variable Residuals, MCR Fitting (Line Plot) 199
Variable Residuals, PCA Fitting (Line Plot) 199
Variances, Individual X-variables (Line Plot) 200
Variances, Individual Y-variables (Line Plot) 200
X-variable Residuals (Line Plot) 201
X-Variance per Sample (Line Plot) 201
X-Variances, One Curve per PC (Line Plot) 202
Y-variable Residuals (Line Plot) 203
Y-Variance Per Sample (Line Plot) 203
Y-Variances, One Curve per PC (Line Plot) 203
2D Scatter Plots 204
Classification Scores (2D Scatter Plot) 204
Cooman's Plot (2D Scatter Plot) 204
Influence Plot, X-variance (2D Scatter Plot) 205
Influence Plot, Y-variance (2D Scatter Plot) 206
Loadings for the X-variables (2D Scatter Plot) 207
Loadings for the Y-variables (2D Scatter Plot) 208
Loadings for the X- and Y-variables (2D Scatter Plot) 209
Loading Weights, X-variables (2D Scatter Plot) 210
Loading Weights, X-variables, and Loadings, Y-variables (2D Scatter Plot) 211
Predicted vs. Measured (2D Scatter Plot) 212
Predicted vs. Reference (2D Scatter Plot) 213
Projected Influence Plot (3 x 2D Scatter Plots) 213
Scatter Effects (2D Scatter Plot) 213
Scores (2D Scatter Plot) 214
Scores and Loadings (Bi-plot) 216
Si vs. Hi (2D Scatter Plot) 218
Si/S0 vs. Hi (2D Scatter Plot) 218
X-Y Relation Outliers (2D Scatter Plot) 219
Y-Residuals vs. Predicted Y (2D Scatter Plot) 220
Y-Residuals vs. Scores (2D Scatter Plot) 222
3D Scatter Plots 222
Influence Plot, X- and Y-variance (3D Scatter Plot) 222
Loadings for the X-variables (3D Scatter Plot) 222
Loadings for the X- and Y-variables (3D Scatter Plot) 222
Loadings for the Y-variables (3D Scatter Plot) 223
Loading Weights, X-variables (3D Scatter Plot) 223
Loading Weights, X- variables, and Loadings, Y-variables (3D Scatter Plot) 223

Scores (3D Scatter Plot) 223
Matrix Plots 224
Leverages (Matrix Plot) 224
Mean (Matrix Plot) 224
Regression Coefficients (Matrix Plot) 225
Response Surface (Matrix Plot) 226
Sample and Variable Residuals, X-variables (Matrix Plot) 227
Sample and Variable Residuals, Y-variables (Matrix Plot) 227
Standard Deviation (Matrix Plot) 227
Cross-Correlation (Matrix Plot) 227
Normal Probability Plots 228
Effects (Normal Probability Plot) 228
Y-residuals (Normal Probability Plot) 229
Table Plots 229
ANOVA Table (Table Plot) 229
Classification Table (Table Plot) 230
Detailed Effects (Table Plot) 231
Effects Overview (Table Plot) 231
Prediction Table (Table Plot) 232
Predicted vs. Measured (Table Plot) 232
Cross-Correlation (Table Plot) 232
Special Plots 232
Interaction Effects (Special Plot) 232
Main Effects (Special Plot) 233
Mean and Standard Deviation (Special Plot) 233
Multiple Comparisons (Special Plot) 234
Percentiles (Special Plot) 234
Predicted with Deviations (Special Plot) 235

Glossary of Terms **237**

Index **269**

What Is New in The Unscrambler 9.6?

For you who have just upgraded your Unscrambler license: here is an overview of the new features since previous versions.

If You Are Upgrading from Version 9.5

These are the first features that were implemented after version 9.5.

Analysis

- Clustering for unsupervised classification of samples. Use menu "**Task - Clustering**"
- Automatic pre-treatments can now be registered in models of reduced size "minimum" and "micro". Access your models from the **Results** menu for registration.

Editor

- Easy filling of missing values in a data table, using either PCA or row column mean analysis. Use menu "**Edit - Fill Missing**" for one-time filling or configure automatic filling using "**File - System Setup**".

Re-formatting and Pre-processing

- Nanometer / Wavenumber unit conversion: two new options in "**Modify - Transform - Spectroscopic**" convert your spectroscopic data from nanometers to wavenumber unit and vice versa.
- Median and Gaussian filtering are two new smoothing options.
- Mean Centering and Standard Deviation scaling are now available as pre-processing. Use new menu option "**Modify - Transform - Center and Scale**".

User-friendliness

- Sample grouping in Editor plots provide group visualization using colors and symbols in line plots, 2D scatter plots,… of raw data. Use menu "**Edit - Options**".
- Remember plot selection and options in saved models. You may now change plots and options in model Viewer. Save the model after those changes. The plots selected on screen prior to saving the model will be displayed again when re-opening the model file.
- Reduce model file size with new format "Micro model". This choice when running a PCA, PCR or PLS saves fewer matrices on file, thus reducing the model file size.

File compatibility

- Improved Excel Import with a new interface for importing from Excel files.
- New import format allows you to import files from Brimrose instruments (BFF3).

Safety

- Lock data set: locked data sets cannot be edited (satisfies the FDA's 21 CRF Part 11 guidelines). Use menu option "**File - Lock**".
- Passwords expire after 70 days (satisfies the FDA's 21 CRF Part 11 guidelines).

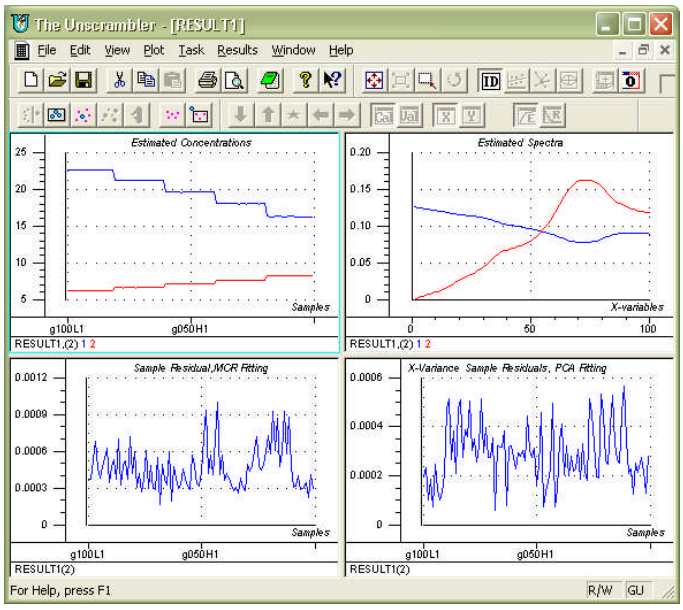
If You Are Upgrading from Version 9.2

These are the first features that were implemented after version 9.2. Look up the previous chapter for newer enhancements.

Analysis

- Multivariate Curve Resolution: resolves mixtures by determining the number of constituents, their profiles and their estimated concentrations. Use menu "**Task - MCR**"

Figure 1 - MCR Overview



Re-formatting and Pre-processing

- Area Normalization , Peak Normalization, Unit Vector Normalization: three new normalization options for pre-processing of multi-channel data.
- Norris Gap derivative, Gap-Segment derivative: two new derivatives implemented in collaboration with Dr. Karl Norris, in replacement for the former "Norris" derivative option.

The former "Norris" derivative from versions 9.2 and earlier will still be supported in auto-pretreatment in The Unscrambler, OLUP and OLUC.

- Savitzky-Golay smoothing and derivatives offer new option settings.

User-friendliness

- File-Duplicate-As 3-D data table: converts an unfolded 2D data table into a 3D format, for modeling with 3-way PLS regression.
- New theoretical chapter introducing Multivariate Curve Resolution, written by Romà Tauler and Anna de Juan.
- New tutorial exercises guiding you through the use of Multivariate Curve Resolution (MCR) modeling.

File compatibility

- Forward compatibility from version 9.0: Read any data or model file built in version 9.x into any other version 9.x. (This does not apply to the new MCR models).
- A new option was introduced when exporting PLS1 models in ASCII format: "Export in the Unscrambler 9.1 format". This ensures maintained compatibility of Unscrambler PLS1 models with Yokogawa analyzers.

New licensing system

- Floating licenses: Define as many user names as you need, and give access to The Unscrambler to a limited number of simultaneous users on your network.
- No delays in receiving Unscrambler upgrades! All license types are available by download.

Plus a number of smaller enhancements.

If You Are Upgrading from Version 9.1

These are the first features that were implemented after version 9.1. Look up the previous chapter for newer enhancements.

Analysis

- Prediction from Three-Way PLS regression models. Open a 3D data table, then use menu "Task-Predict".

Re-formatting and Pre-processing

- Find/replace functionality in the Editor
- Extended Multiplicative Scatter Correction (EMSC)
- Standard Normal Variate (SNV)

Visualisation

- Two new plots are available for Analysis of Effects results: "Main effects" and "Interaction effects".

- Correlation matrix directly available as a matrix plot in Statistics results.
- Easy sample and variable identification on line plots.

Compatibility with other software

- Compatibility with databases: Oracle, MySQL, MS Access, SQL Server 7.0, ODBC.
- User-Defined Import (UDI): Import any file format into The Unscrambler!

Plus various smaller enhancements and bug fixes.

If You Are Upgrading from Version 8.0.5

These are the first features that were implemented after version 8.0.5. Look up the previous chapters for newer enhancements.

Analysis

- New analysis method: Three-Way PLS regression. Open a 3D data table, then use menu “Task-Regression”.
  The following key features can be named: Two validation methods available (Cross-Validation and Test Set), Scaling and Centering options, over 50 pre-defined plots to view the model results, over 60 importable result matrices.
- The following data pretreatments and their combinations are available as automatic pretreatments in Classification and Prediction: Smoothing, Normalize, Spectroscopic, MSC, Noise, Derivatives, Baselines. Combinations of these pretreatments are also supported in auto-pretreatments.

3D Editor

- Toggle between the 12 possible layouts of 3D tables with submenus in the Modify menu or using Ctrl+3
- Create Primary Variable and Secondary Variable sets for use in 3-Way analysis. Use menu “Modify-Edit Set” on an active 3D table.

User-friendliness

- Optimized PC-Navigation toolbar. Freely switch PC numbers by a simple click on the “Next horizontal PC”, “Previous horizontal PC”, “Next vertical PC”, “Previous vertical PC” and “Suggested PC” buttons, or use the corresponding arrow keys on your keyboard. The PC-Navigation tool is available on all PCA, PCR, PLS-R and Prediction result plots.
- A shortcut key Ctrl+R was created for “File-Import-Unscrambler Results”

Compatibility with other software

- Importation of 3D tables from Matlab supported. Use menu “File-Import 3D-Matlab”
- Importation of *.F3D file format from Hitachi supported. Use menu “File-Import 3D-F3D”
- Importation of files from Analytical Spectral Devices software supported (file extensions: *.001 and *.asd). Use menu “File-Import-Indico”

Visualisation

- Passified variables are displayed in a different color from non-passified variables on Bi-Plots, so that they are easily identified.
- Plot header and axes denomination are shown on 2D Scatter plots, 3D Scatter plots, histogram plots, Normal probability plots and matrix plots of raw data.

Plus several bug fixes and minor improvements.

If You Are Upgrading from Version 8.0

These are the first features that were implemented after version 8.0. Look up the previous chapters for newer enhancements.

Analysis

- In SIMCA-classification results, significance level "None" was introduced in Si vs Hi and Si/S0 vs Hi plots. This option allows to display these plots with no significance limits, as was implemented for Coomans'plot in version 8.0
- The chosen variable weights are more accurately indicated than in previous versions in the PCA and Regression dialogs
- Weighting is free for each model term, except with the Passify option which automatically passifies all interactions and squares of passified main effects. The user can change this default by using the "Weights..." button in the PCA and Regression dialogs.

Visualisation

- Passified variables are displayed in a different color from non-passified variables on Loadings and Correlation Loadings plots so that they are easily identified.
- When computing a PCR or PLS-R model with Uncertainty Test, the significant X-variables are marked by default when opening the results Viewer

Compatibility with other software

- Importation of file formats *.asc, *.scn and *.autoscan from Guided Wave is now supported (CLASS-PA and SpectrOn software)
- Importing very large ASCII data files is subsequently faster than in previous versions

Plus several bug fixes and minor improvements.

If You Are Upgrading from Version 7.8

These are the first features that were implemented after version 7.8. Look up the previous chapters for newer enhancements.

User-friendliness

- Undo-Redo buttons are available for most Editor functions.
- A Guided Expression dialog makes the Compute function simpler and more intuitive to use.
- Sort Variable Sets and Sort Sample Sets are now available even in the presence of overlapping sets.
- Switch PC numbers by a simple click on the “Next PC” and “Previous PC” buttons in most plots of the PCA, PCR and PLS regression results.
- New function in the marking toolbar: Reverse marking
- Possibility to save plots in five image formats (Bitmap, Jpeg, Gif, Portable network graphics and TIFF)
- An « Undo Adjust » button allows you to regret forcing a simplex onto your mixture design
- New User Guide documentation in html format – click and read!

Visualisation

- Sample grouping options let you choose how many groups to use, which sample ID should be displayed on the plot and how many decimals/characters to be displayed
- Possibility to perform Sample Grouping with symbols instead of colours. It allows to visualise groups also when printing plots in black & white
- The Loadings plot replaces the Loading Weights plot in Regression Overview results, thus allowing easy access to the Correlation loadings plot.
- Select « None » as significance limits in Cooman’s plot (classification)

Analysis

- Improved Passify weights
- Improved Uncertainty test (Jack-knife variance estimates)
- The raw regression coefficients are available through the Plot menu. In addition, B0 or B0W values are indicated on the regression coefficients plots
- Skewness is included in the View-Statistics tables

Traceability

- Data and model files information indicate the software version that was used to create the file.
- The Empty button in File-Properties-Log can be disabled in the administrator system setup options, preventing the user from deleting the log of performed operations.

If You Are Upgrading from Version 7.6

These are the first features that were implemented after version 7.6. Look up the previous chapters for newer enhancements.

Easy and automated import of ASCII files:

You can launch The Unscrambler from an external application and automatically read the contents of ASCII files into a new Unscrambler data table.

Enhanced Import features:

Space is no longer a default item delimiter when importing from ASCII files. Instead it is available as an option among other delimiters.

Enhanced Editor functions:

1. You may now Reverse Sample Order or Reverse Variable Order in your data table. It is also possible to Sort by Sample Sets or by Variable Sets.
2. It is now possible to create new Sample Sets from a Category Variable.
3. Sample and Variable Sets now support any Set size, even if the range is non-continuous.

Improved Recalculate options:

1. You may now Passify X- or Y-variables when recalculating your PCA, PCR or PLS model. The variables are kept in the analysis but are weighted close to zero so as not to influence the model.
2. A bug fix allows you to keep out Y-variables by using “Recalculate Without Marked”.

Improved D-optimal design interface:

1. More user-friendly definition of multi-linear constraints.
2. Better information about the condition number of your design.

New function User Defined Analysis:

You may now add your own analysis routines for 3D data. This works with DLLs, in the same way as User Defined Transformations.

If You Are Upgrading from Version 7.5

These are the first features that were implemented after version 7.5. Look up the previous chapters for newer enhancements.

New data structure:

It is now possible to import or convert data into a 3-D structure.

Work with category variables:

Easier importation of category variables.

Customizable model size:

Save your models in the appropriate size: Full, Compact or Minimum.

Loadings:

Correlation Loadings are now implemented and help you interpret variable correlations in Loading plots.

Export to and Import from Matlab:

You can directly export data to Matlab, or import data from Matlab including sample and variable names.

New import format:

MVACDF.

If You Are Upgrading from Version 7.01

These are the first features that were implemented after version 7.01. Look up the previous chapters for newer enhancements.

Martens' Uncertainty test:

New and unique method based on "Jack-knifing", for safer interpretation with significance testing. The new method developed by Dr. Harald Martens shows you which variables are significant or not, the uncertainty estimates for the variables and the model robustness.

New experimental plans:

Mixtures, D-optimal designs and combination of those. Analysis with PLS or Response Surface.

Live 3D rotation of scatter plots:

Get a visual understanding of the structure of your data through real-time 3D rotation. Applies to 3D-scatter plots, matrix plots and response surface plots.

More professional presentation of your results:

To ease your documentation work, new gray-tone schemes and features were added to separate information also on black & white printouts.

Add your own transformation routines:

The Unscrambler can now utilize transformation DLLs so you can use your favorite pre-processing methods that you develop yourself or get from algorithm libraries. At prediction and classification of new data, The Unscrambler applies all pre-processing stored with the model.

Easier to detect outliers:

Hotelling T2 statistics allow outlier boundaries to be visualized as ellipses in your score plots, and make the interpretation very simple.

Import of Excel 97 files:

Import of Excel 97 files with named ranges and embedded charts now fully supported.

Recalculation is now possible after all analyses:

Recalculation now also works for Analysis of Effects and Response Surface.

Print plots from several windows simultaneously:

A new print dialog for viewer documents makes it possible to print all visible plots on screen (2 or 4) on the same sheet of paper.

Level markers in contour plots:

In contour plots, level markers on contour lines are now implemented.

New added matrix when exporting:

Extended export model to ASCII-MOD format. If exporting full PCA or full Regression model, the matrix "Tai" is included on the output ASCII-MOD file as the last model matrix, but before any MSC model matrix.

What is The Unscrambler?

A brief review of the tasks that can be carried out using The Unscrambler.

The main purpose of The Unscrambler is to provide you with tools which can help you analyze multivariate data. By this we mean finding variations, co-variations and other internal relationships in data matrices (tables). You can also use The Unscrambler to design the experiments you need to perform to achieve results which you can analyze.

The following are the basic types of problems that can be solved using The Unscrambler:

- Design experiments, analyze effects and find optima;
- Re-format and pre-process your data to enhance future analyses;
- Find relevant variation in one data matrix;
- Find relationships between two data matrices (X and Y);
- Validate your multivariate models with Uncertainty Testing;
- Resolve unknown mixtures by finding the number of pure components and estimating their concentration profiles and spectra;
- Find relationships between one response data matrix (Y) and a “cube” of predictors (three-way data X);
- Predict the unknown values of a response variable;
- Classify unknown samples into various possible categories.

You should always remember, however, that there is no point in trying to analyze data if they do not contain any meaningful information. Experimental design is a valuable tool for building data tables which give you such meaningful information. The Unscrambler can help you do this in an elegant way.

The Unscrambler® satisfies the FDA's requirements for 21 CFR Part 11 compliance.

Make Well-Designed Experimental Plans

Choosing your samples carefully increases the chance of extracting useful information from your data. Furthermore, being able to actively experiment with the variables also increases the chance. The critical part is deciding which variables to change, which intervals to use for this variation, and the pattern of the experimental points.

The purpose of ***experimental design*** is to generate experimental data that enable you to find out which ***design variables*** (X) have an influence on the ***response variables*** (Y), in order to understand the interactions between the design variables and thus determine the optimum conditions. Of course, it is equally important to do this with a minimum number of experiments to reduce costs. An experimental design program should offer appropriate design methods and encourage good experimental practice, i.e. allow you to perform few but useful experiments which span the important variations.

Screening designs (e.g. fractional, full factorial and Plackett-Burman) are used to find out which design variables have an ***effect*** on the responses, and are suitable for collection of data spanning all important variations.

Optimization designs (e.g. central composite, Box-Behnken) aim to find the optimum conditions for a process and generate non-linear (quadratic) models. They generate data tables that describe relationships in more detail, and are usually used to refine a model, i.e. after the initial screening has been performed.

Whether your purpose is screening or optimization, there may be ***multi-linear constraints*** among some of your design variables. In such a case you will need a **D-optimal design**.

Another special case is that of **mixture designs**, where your main design variables are the ***components of a mixture***. The Unscrambler provides you with the classical types of mixture designs, with or without additional constraints.

There are several methods for analysis of experimental designs. The Unscrambler uses Analysis Of Effects (ANOVA) and MLR as its default methods for ***orthogonal*** designs (i.e. not mixture or D-optimal), but you can also use other methods, such as PCR or PLS.

Reformat, Transform and Plot your Data

Raw data may have a distribution that is not optimal for analysis. Background effects, measurements in different units, different variances in variables etc. may make it difficult for the methods to extract meaningful information. ***Preprocessing*** reduces the "noise" introduced by such effects.

Before you even reach that stage, you may need to look at your data from a slightly different point of view. Sorting samples or variables, transposing your data table, changing the layout of a 3D data table are examples of such ***re-formatting*** operations.

Whether your data have been re-formatted and pre-processed or not, a quick ***plot*** may tell you much more than is to be seen with the naked eye on a mere collection of numbers. Various types of plots are available in the Unscrambler, they help you visually check individual variable distributions, study the correlation among two variables or examine your samples as for example a 3-dimensional swarm of points or a 3-D landscape.

Study Variations among One Group of Variables

A common problem is to determine which variables actually contribute to the variation seen in a given data matrix; i.e. to find answers to questions such as

- "Which variables are necessary to describe the samples adequately?";
- "Which samples are similar to each other?";
- "Are there groups of samples in my data?";
- "What is the meaning of these sample patterns?".

The Unscrambler finds this information by decomposing the data matrix into a structure part and a noise part, using a technique called ***Principal Component Analysis (PCA)***.

Other Methods to Describe One Group of Variables

Classical ***descriptive statistics*** are also available in The Unscrambler. Mean, standard deviation, minimum, maximum, median and quartiles provide an overview of the univariate distributions of your variables, allowing for comparisons between variables. In addition, the ***correlation*** matrix provides a crude summary of the co-variations among variables.

In the case of instrumental measurements (such as spectra or voltammograms) performed on samples representing mixtures of a few pure components at varying concentrations or at different stage of a process (such as chromatography), the Unscrambler offers a method for recovering the unknown concentrations, called ***Multivariate Curve Resolution (MCR)***.

Study Relations between Two Groups of Variables

Another common problem is establishing a ***regression model*** between two data matrices. For example, you may have a lot of inexpensive measurements (X) of properties of a set of different solutions, and want to relate these measurements to the concentration of a particular compound (Y) in the solution, found by a reference method.

In order to do this, we have to find the relationship between the two data matrices. This task varies somewhat depending on whether the data has been generated using statistical experimental design (i.e. designed data) or has simply been collected, more or less at random, from a given population (i.e. non-designed data).

How to Analyze Designed Data Matrices

The variables in designed data tables (excluding mixture or D-optimal designs) are orthogonal. Traditional statistical methods such as ANOVA and MLR are well suited to make a regression model from orthogonal data tables.

How to Analyze Non-designed Data Matrices

The variables in non-designed data matrices are seldom orthogonal, but rather more or less collinear with each other. MLR will most likely fail in such circumstances, so the use of ***projection techniques*** such as Principal Component Regression (PCR) or Partial Least Squares (PLS) is recommended.

Validate your Multivariate Models with Uncertainty Testing

Whatever your purpose in multivariate modelling – explore, describe precisely, build a predictive model – ***validation*** is an important issue. Only a proper validation can ensure that your results are not too highly dependent on some extreme samples, and that the predictive power of your regression model meets your expectations.

With the help of ***Martens' Uncertainty Test***, the power of cross validation is further increased and allows you to

- Study the influence of individual samples on your model on powerful, simple to interpret graphical representations;
- Test the significance of your predictor variables and remove unimportant predictors from your PLS or PCR model.

Make Calibration Models for Three-way Data

Regression models are also relevant for data which do not fit in a two-dimensional matrix structure. However, three-way data require a specific method because the usual vector / matrix calculations no longer apply.

Three-way PLS (or tri-PLS) takes the principles of PLS further and allows you to build a regression model which explains the variations in one or several responses (Y-variables) to those of a 3-D array of predictor variables, structured as Primary and Secondary X-variables (or X1- and X2-variables).

Estimate New, Unknown Response Values

A regression model can be used to ***predict*** new, i.e. unknown, Y-values. Prediction is a useful technique as it can replace costly and time consuming measurements. A typical example is the prediction of concentrations from absorbance spectra instead of direct measurements of them.

Classify Unknown Samples

Classification simply means to find out whether new samples are similar to classes of samples that have been used to make models in the past. If a new sample fits a particular model well, it is said to be a member of that class.

Many analytical tasks fall into this category. For example, raw materials may be sorted into “good” and “bad” quality, finished products classified into grades “A”, “B”, “C”, etc.

Reveal Groups of Samples

Clustering is an attempt to group samples into ‘k’ clusters based on specific distance measurements.

In The Unscrambler, you may apply clustering on your data, using the K-Means algorithm. Seven different types of distance measurements are provided with the algorithm.

Data Collection and Experimental Design

In this chapter, you may read about all the aspects of data collection covered in The Unscrambler:

- How to collect "good" data for a future analysis, with special emphasis given to experimental design methods;
- Specific issues related to three-way data;
- How data entry and experimental design generation are taken care of in practice in The Unscrambler.

Principles of Data Collection and Experimental Design

Learn how to generate the experimental data that will be best suited for the problems you want to solve or the questions you want to explore.

Data Collection Strategies

The aim of multivariate data analysis is to extract information from a data table. The data can be collected from various sources or designed with a specific purpose in mind.

When collecting new data for multivariate modeling, you should usually pay attention to the following criteria:

- ***Efficiency*** - get more information from fewer experiments;
- ***Focusing*** - collect only the information you really need.

There are four basic ways to collect data for an analysis:

- Get hold of ***historical*** data (from a database, from plant records, etc.);
- Collect ***new*** data: record measurements directly from the production line, make observations in the fish farms, etc… This will ensure that the data apply to the system that you are studying, today (not another system, three years ago);
- Make your own ***experiments*** by disturbing the system you are studying. Thus the data will encompass more variation than is to be seen in a stable system running as usual.
- ***Design*** your experiments in a structured, mathematical way. By choosing symmetrical ranges of variation and applying this variation in a balanced way among the variables you are studying, you will end up with data where effects can be studied in a simple and powerful way. You will also have better possibilities of testing the significance of the effects and the relevance of the whole model.

Experimental design is a useful complement to multivariate data analysis because it generates "structured" data tables, i.e. data tables that contain an important amount of ***structured variation***. This underlying structure will then be used as a basis for multivariate modeling, which will guarantee stable and robust model results.

More generally, a careful sample selection increases the chances of extracting useful information from your data. When you have possibilities to actively perturb your system (experiment with the variables) these chances become even bigger. The critical part is to decide which variables to change, the intervals for this variation, and the pattern of the experimental points.

What Is Experimental Design?

Experimental design is a strategy to gather ***empirical*** knowledge, i.e. knowledge based on the analysis of experimental data and not on theoretical models. It can be applied whenever you intend to investigate a phenomenon in order to gain understanding or improve performance.

Building a design means carefully choosing a small number of experiments that are to be performed under controlled conditions. There are four interrelated steps in building a design:

1. Define an ***objective*** to the investigation, e.g. "better understand" or "sort out important variables" or "find optimum".
4. Define the variables that will be controlled during the experiment (***design variables***), and their levels or ranges of variation.
5. Define the variables that will be measured to describe the outcome of the experimental runs (***response variables***), and examine their precision.
6. Choose among the available ***standard designs*** the one that is compatible with the objective, number of design variables and precision of measurements, and has a reasonable cost.

Standard designs are well-known classes of experimental designs which can be generated automatically in The Unscrambler as soon as you have decided on the objective, the number and nature of design variables, the nature of the responses and the number of experimental runs you can afford. Generating such a design will provide you with the list of all experiments you must perform to gather enough information for your purposes.

Various Types of Variables in Experimental Design

This section introduces the nomenclature of variable types used in The Unscrambler. Most of these names are commonly used in the standard literature on experimental design; however the use made of these names in The Unscrambler may be somewhat different from what you are expecting. Therefore we recommend that you read this section before proceeding to more details about the various types of designs.

Design Variables

Performing designed experiments is based on controlling the variations of the variables for which you want to study the effects. Such variables with controlled variations are called **design variables**. They are sometimes also referred to as factors.

In The Unscrambler, a design variable is completely defined by:

- Its name;
- Its type: continuous or category;
- Its levels.

Note: in some cases (D-optimal or Mixture designs), the variables with controlled variations will be referred to using other names: "mixture variables" or "process variables". Read more in Designs for Simple Mixture Situations, D-Optimal Designs Without Mixture Variables and D-Optimal Designs With Mixture Variables.

Continuous Variables

All variables that have numerical values and that can be measured quantitatively are called **continuous variables**. This may be somewhat abusive in the case of discrete quantitative variables, such as counts. It reflects the implicit use which is made of these variables, namely the modeling of their variations using continuous functions.

Examples of ***continuous variables*** are: temperature, concentrations of ingredients (e.g. in %), pH, length (e.g. in mm), age (e.g. in years), number of failures in one year, etc.

Levels of Continuous Variables

The variations of continuous design variables are usually set within a predefined range, which goes from a ***lower level*** to an ***upper level***. Those two levels have to be specified when defining a continuous design variable. You can also choose to specify more levels between the extremes if you wish to study some values specifically.

If only two levels are specified, the other necessary levels will be computed automatically. This applies to ***center samples*** (which use a mid-level, half-way between lower and upper), and ***star samples*** in optimization designs (which use extreme levels outside the predefined range). See sections Center Samples and Sample Types in Central Composite Designs for more information about center and star samples.

Note: If you have specified more than two levels, center samples will not be computed.

Category Variables

In The Unscrambler, all non-continuous variables are called **category variables**. Their levels can be named, but not measured quantitatively.

Examples of ***category variables*** are: color (Blue, Red, Green), type of catalyst (A, B, C, D), place of origin (Africa, The Caribbeans)…

Binary variables are a special type of category variables. They have only two levels and symbolize an alternative.

Examples of ***binary variables*** are: use of a catalyst (Yes/No), recipe (New/Old), type of electric power (AC/DC), type of sweetener (Artificial/ Natural)...

Levels of Category Variables

For each category variable, you have to **specify all levels**.

Note: Since there is a kind of quantum jump from one level to another (there is no intermediate level in-between), you cannot directly define ***center samples*** when there are category variables.

Non-design Variables

In The Unscrambler, all variables appearing in the context of designed experiments which are not themselves design variables, are called **non-design variables**.

This is generally synonymous to ***Response variables***, i.e. measured output variables that describe the outcome of the experiments.

Mixture Variables

If you are performing experiments where some ingredients have to be mixed according to a recipe, you may be in a situation where the amounts of the various ingredients cannot be varied independently from each other. In such a case, you will need to use a special kind of design called Mixture design, and the variables with "controlled" variations are then called **mixture variables**.

An example of a ***mixture situation*** is blending concrete from the following three ingredients: cement, sand and water. If you increase the percentage of water in the blend with 10%, you will have to reduce the proportions of one of the other ingredients (or both) so that the blend still amounts to 100%.

However, there are many situations where ingredients are blended, which do not require a mixture design. For instance in a water solution of four ingredients whose proportions do not exceed a few percent, you may vary the four ingredients independently from each other and just add water at the end as a "filler". Therefore you will have to think carefully before deciding whether you own recipe requires a mixture design or not!

Read more about Mixture designs in chapter Designs for Simple Mixture Situations p.30.

Process Variables

In a mixture situation, you may also want to investigate the effects of variations in some other variables which are not themselves a component of the mixture. Such variables are called **process variables** in The Unscrambler.

Typical ***process variables*** are: temperature, stirring rate, type of solvent, amount of catalyst, etc.

The term ***process variables*** will also be used for non-mixture variables in a design dealing with variables that are linked by Multi-Linear Constraints (D-Optimal design). Read more about D-Optimal designs in chapter Introduction to the D-Optimal Principle p.35.

Investigation Stages and Design Objectives

Depending on the stage of the investigation, the amount of information you wish to collect, and the resources that are available to achieve your goal, you will have to choose an adequate design among those available in The Unscrambler. These are the most common standard designs, dealing with several continuous or category variables that can be varied independently of each other, as well as mixture or D-optimal designs.

Screening

When you start a new investigation or a new product development, there is usually a large number of potentially important variables. At this stage, the aim of the experiments is to find out which are the ***most important variables***. This is achieved by including many variables in the design, and roughly estimating the effect of each design variable on the responses with the help of a screening design. The variables which have "large" effects can be considered as important.

Main Effects and Interactions

The variation in a response generated by varying a design variable from its low to its high level is called the ***main effect*** of that design variable on that response. It is computed as the ***linear*** effect of the design variable over its whole range of variation. There are several ways to judge the importance of a main effect, for instance significance testing or use of a normal probability plot of effects.

Some variables can be considered important even though they do not have an important impact on a response by themselves. The reason is that they can also be involved in an ***interaction***. There is an interaction between two variables when changing the level of one of those variables modifies the effect of the second variable on the response.

Interaction effects are computed using the products of several variables. There can be various orders of interaction: two-factor interactions involve two design variables, three-factor interactions involve three of them, and so on. The importance of an interaction can be assessed with the same tools as for main effects.

Design variables that have an important main effect are important variables. Variables that participate in an important interaction, even if their main effects are negligible, are also important variables.

Models for Screening Designs

Depending on how precisely you want to screen the potentially influent variables and describe how they affect the responses, you have to choose the adequate shape of the model that relates response variations to design variable variations. The Unscrambler contains two standard choices:

- The simplest shape is a ***linear model***. If you choose a linear model, you will investigate main effects only;
- If you are also interested in the possible interactions between several design variables, you will have to include ***interaction effects*** in your model in addition to the linear effects.

When building a mixture or D-optimal design, you will need to choose a model shape explicitly, because the adequate type of design depends on this choice. For other types of designs, the model choice is implicit in the design you have selected.

Optimization

At a later stage of investigation, when you already know which variables are important, you may wish to study the effects of a few major variables in more detail. Such a purpose will be referred to as ***optimization***. Another term often used for this procedure, especially at the analysis stage, is ***response surface modeling***.

Objectives for Optimization

Optimization designs actually cover quite a wide range of objectives. They are particularly useful in the following cases:

- ***Maximizing a single response***, i.e. to find out which combinations of design variable values lead to the maximum value of a specific response, and how high this maximum is.
- ***Minimizing a single response***, i.e. to find out which combinations of design variable values lead to the minimum value of a specific response, and how low this minimum is.
- ***Finding a stable region***, i.e. to find out which combinations of design variable values lead closely enough to the target value of a specific response, while a small deviation from those settings would cause negligible change in the response value.
- ***Finding a compromise between several responses***, i.e. to find out which combinations of design variable values lead to the best compromise between several responses.
- ***Describing response variations***, i.e. to model response variations inside the experimental region as precisely as possible in order to predict what will happen if the settings of some design variables have to be changed in the future.

Models for Optimization Designs

The underlying idea for optimization designs is that the model should be able to describe a response surface which has a minimum or a maximum inside the experimental range. To achieve that purpose, linear and interaction effects are not sufficient. This is why an optimization model should also include ***quadratic effects***, i.e. square effects, which describe the concavity or convexity of a surface.

A model that includes linear, interaction and quadratic effects is called a ***quadratic model***.

Designs for Unconstrained Screening Situations

The Unscrambler provides three classical types of screening designs for unconstrained situations:

- ***Full factorial designs*** for any number of design variables between 2 and 6; the design variables may be continuous or category, with 2 to 20 levels each.
- ***Fractional factorial designs*** for any number of 2-level design variables (continuous or category) between 3 and 15.
- ***Plackett-Burman designs*** for any number of 2-level design variables (continuous or category) between 4 and 32.

Full Factorial Designs

Full factorial designs combine all defined levels of all design variables. For instance, a full factorial design investigating one 2-level continuous variable, one 3-level continuous variable and one 4-level category variable will include 2x3x4=24 experiments.

Among other properties, full factorial designs are perfectly balanced, i.e. each level of each design variable is studied an equal number of times in combination with each level of each other design variable.

Full factorial designs include enough experiments to allow use of a model with all interactions. Thus, they are a logical choice if you intend to study interactions in addition to main effects.

Fractional Factorial Designs

In the specific case where you have only 2-level variables (continuous with lower and upper levels, and/or binary variables), you can define fractions of full factorial designs that enable you to investigate as many design variables as full factorial designs with fewer experiments. These "cheaper" designs are called ***fractional factorial designs***.

Given that you already have a full factorial design, the most natural way to build a fractional design is to use only half the experimental runs of the original design. For instance, you might try to study the effects of three design variables with only 4 (2^2) instead of 8 (2^3) experiments. Larger factorial designs admit fractional designs with a higher degree of fractionality, i.e. even more economical designs, such as investigating nine design variables with only 16 (2^4) experiments instead of 512 (2^9). Such a design can be referred to as a 2^{9-5} design; its degree of fractionality is 5. This means that you investigate nine variables at the usual cost of four (thus saving the cost of five).

Example of a Fractional Factorial Design

In order to better understand the principles of fractionality, let us illustrate how a fractional factorial is built in the following concrete case: computing the half-fraction of a full factorial with four variables (2^{4-1}).

In the following tables, the design variables are named A, B, C, D, and their lower and upper levels are coded – and +, respectively.

First, let us build a full factorial design with only variables A, B, C (2^3), as seen below:

Full factorial design 2^3

Experiment	A	B	C
1	–	–	–
2	–	–	+
3	–	+	–
4	–	+	+
5	+	–	–
6	+	–	+
7	+	+	–
8	+	+	+

If we now build additional columns, computed from products of the original three columns A, B, C, we get the new table shown hereafter. These additional columns will symbolize the interactions between the design variables.

Full factorial design 2^3 *with interaction columns*

Experiment	A	B	C	AB	AC	BC	ABC
1	–	–	–	+	+	+	–
2	–	–	+	+	–	–	+
3	–	+	–	–	+	–	+
4	–	+	+	–	–	+	–
5	+	–	–	–	–	+	+

6	+	–	+	–	+	–	–
7	+	+	–	+	–	–	–
8	+	+	+	+	+	+	+

We can see that none of the seven columns are equal; this means that the effects symbolized by these columns can all be studied independently of each other, using only 8 experiments.

If we now use the last column to study the main effect of an additional variable, D, instead of ABC:

Fractional factorial design 2^{4-1}

Experiment	A	B	C	D
	–	–	–	–
	–	–	+	+
	–	+	–	+
	–	+	+	–
	+	–	–	+
	+	–	+	–
	+	+	–	–
	+	+	+	+

It is obvious that the new design allows the main effects of the 4 design variables to be studied independently of each other; but what about their interactions? Let us try to build all 2-factor interaction columns, illustrated in the table hereafter. Since only seven different columns can be built out of 8 experiments (except for columns with opposite signs, which are not independent), we end up with the following table:

Fractional factorial design 2^{4-1} *with interaction columns*

Experiment	A	B	C	D	AB = CD	AC = BD	BC = AD
1	–	–	–	–	+	+	+
2	–	–	+	+	+	–	–
3	–	+	–	+	–	+	–
4	–	+	+	–	–	–	+
5	+	–	–	+	–	–	+
6	+	–	+	–	–	+	–
7	+	+	–	–	+	–	–
8	+	+	+	+	+	+	+

As you can see, each of the last three columns is common to two different interactions (for instance, AB and CD share the same column).

Confounding

Unfortunately, as the example shows, there is a price to be paid for saving on the experimental costs! If you invest less, you will also harvest less...

In the case of fractional factorials, this means that if you do not use the full factorial set of experiments, you might not be able to study the interactions as well as the main effects of all design variables. This happens because of the way those fractions are built, using some of the resources that would otherwise have been devoted to the study of interactions, merely to study main effects of more variables instead.

This side effect of some fractional designs is called **confounding**. Confounding means that some effects cannot be studied independently of each other.

For instance, in the above example, the 2-factor interactions are confounded with each other. The practical consequences are the following:

- All main effects can be studied independently of each other, and independently of the interactions;
- If you are interested in the interactions themselves, using this specific design will only enable you to detect whether some of them are important. You will not be able to decide which are the important ones. For instance, if AB (confounded with CD, "AB=CD") turns out as significant, you will not know whether AB or CD (or a combination of both) is responsible for the observed effect.

The list of confounded effects is called the **confounding pattern** of the design.

Resolution of a Fractional Design

How well a fractional factorial design avoids confounding is expressed through its **resolution**. The three most common cases are as follows:

- ***Resolution III designs***: Main effects are confounded with 2-factor interactions.
- ***Resolution IV designs***: Main effects are free of confounding with 2-factor interactions, but 2-factor interactions are confounded with each other.
- ***Resolution V designs***: Main effects and 2-factor interactions are free of confounding.

Definition: In a **Resolution *R*** design, effects of order *k* are free of confounding with all effects of order less than *R-k*.

In practice, before deciding on a particular factorial design, check its resolution and its confounding pattern to make sure that it fits your objectives!

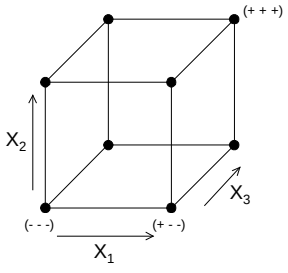
Plackett-Burman Designs

If you are interested in main effects only, and if you have many design variables to investigate (let us say more than 10), Plackett-Burman designs may be the solution you need. They are very economical, since they require only 1 to 4 more experiments than the number of design variables.

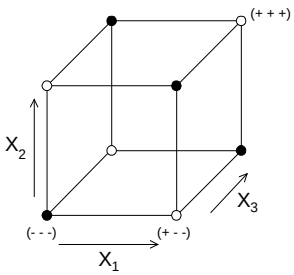
Examples of Factorial Designs

A screening situation with three design variables:

Screening design; three design variables



Full factorial 2^3



Fractional factorial 2^{3-1}

Designs for Unconstrained Optimization Situations

The Unscrambler provides two classical types of optimization designs:

- ***Central Composite*** designs for 2 to 6 continuous design variables;
- ***Box-Behnken*** designs for 3 to 6 continuous design variables.

Note: Full factorial designs with 3-level (or more) continuous variables can also be used as optimization designs, since the number of levels is compatible with a quadratic model. They will not be described any further here.

Central Composite Designs

Central composite designs (CCD) are extensions of 2-level full factorial designs which enable a quadratic model to be fitted by including more levels in addition to the specified lower and upper levels.

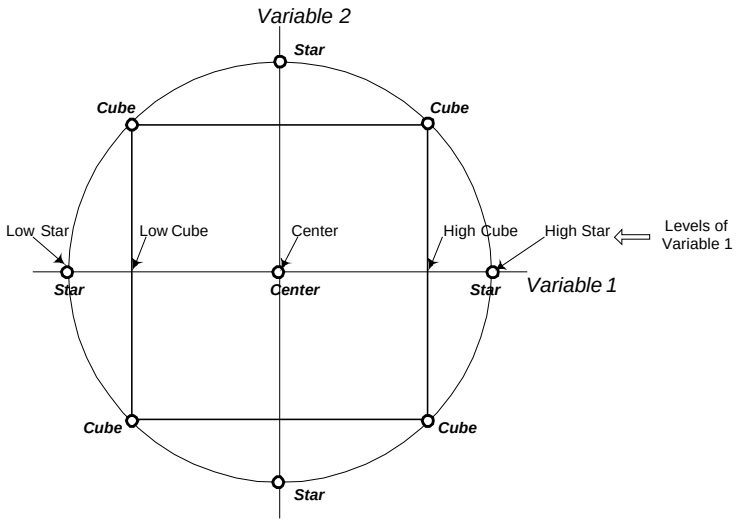
A central composite design consists of three types of experiments:

- **Cube samples** are experiments which cross lower and upper levels of the design variables; they are the "factorial" part of the design;
- **Center samples** are the replicates of the experiment which cross the mid-levels of all design variables; they are the "inside" part of the design.
- **Star samples** are used in experiments which cross the mid-levels of all design variables except one with the extreme (star) levels of the last variable. Those samples are specific to central composite designs.

Properties of a Central Composite Design

Let us illustrate this with a simple example: a CCD with two design variables.

Central composite design with two design variables



As you can see, each design variable has 5 levels: Low Star, Low Cube, Center, High Cube, High Star. Low Cube and High Cube are the lower and upper levels that you specify when defining the design variable.

- The four cube samples are located at the corners of a square (or a cube if you have 3 variables, or a hyper-cube if you have more), hence their name;
- The center samples are located at the center of the square;

- The four star samples are located outside the square; by default, their distance to the center is the same as the distance from the cube samples to the center, i.e. here:

$$\sqrt{2} \cdot \frac{\left(High\ Cube - Low\ Cube\right)}{2}$$

As a result, all cube and star samples are located on the same circle (or sphere if you have 3 design variables). From that fact follows that all cube and star samples will have the same leverage, i.e. the information they carry will have equal weight on the analysis. This property, called ***rotatability***, is important if you want to achieve uniform quality of prediction in all directions from the center.

However, if for some reason those levels are impossible to achieve in the experiments, you can tune the "star distance to center" factor down to a minimum of 1. Then the star points will lie at the center of the cube faces.

Another way to keep all experiments within a manageable range when the default star levels are too extreme, is to use the optimal star sample distance, but shrink the high and low cube levels. This will result in a smaller investigated range, but will guarantee a rotatable design.

Box-Behnken Designs

Box-Behnken designs are not built on a factorial basis, but they are nevertheless good optimization designs with simple properties.

In a Box-Behnken design, all design variables have exactly three levels: Low Cube, Center, High Cube. Each experiment crosses the extreme levels of 2 or 3 design variables with the mid-levels of the others. In addition, the design includes a number of center samples.

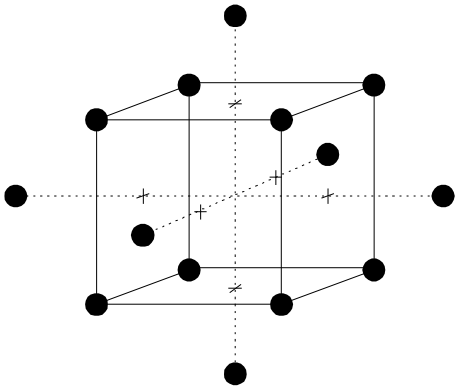
The properties of Box-Behnken designs are the following:

- The actual range of each design variable is Low Cube to High Cube, which makes it easy to handle;
- All non-center samples are located on a sphere, thus achieving rotatability.

Examples of Optimization Designs

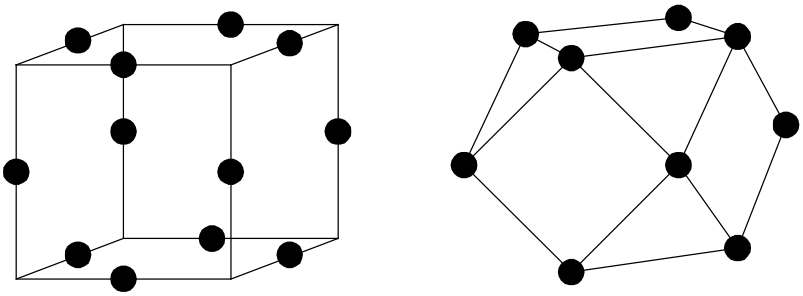
A central composite design for three design variables:

Central composite design; three design variables



In the figure below, the Box-Behnken design is shown drawn in two different ways. In the left drawing you see how it is built, while the drawing to the right shows how the design is rotatable.

Box-Behnken design



Designs for Constrained Situations, General Principles

This chapter introduces "tricky" situations in which classical designs based upon the factorial principle do not apply. Here, you will learn about two specific cases:

1. Constraints between the levels of several design variables;
2. A special case: mixture situations.

Each of these situations will then be described extensively in the next chapters.

Note: To understand the sections that follow, you need basic knowledge about the purposes and principles of experimental design. If you have never worked with experimental design before, we strongly recommend that you read about it in the previous sections (see What Is Experimental Design?) before proceeding with this chapter.

Constraints Between the Levels of Several Design Variables

A manufacturer of prepared foods wants to investigate the impact of several processing parameters on the sensory properties of cooked, marinated meat. The meat is to be first immersed in a marinade, then steam-cooked, and finally deep-fried. The steaming and frying temperatures are fixed; the marinating and cooking times are the process parameters of interest.

The process engineer wants to investigate the effect of the three process variables within the following ranges of variation:

Ranges of the process variables for the cooked meat design

Process variable	Low	High
Marinating time	6 hours	18 hours
Steaming time	5 min	15 min
Frying time	5 min	15 min

A full factorial design would lead to the following "cube" experiments:

The cooked meat full factorial design

Sample	Mar. Time	Steam. Time	Fry. Time
1	6	5	5
2	18	5	5
3	6	15	5
4	18	15	5

5	6	5	15
6	18	5	15
7	6	15	15
8	18	15	15

When seeing this table, the process engineer expresses strong doubts that experimental design can be of any help to him. “Why?” asks the statistician in charge. “Well,” replies the engineer, “if the meat is steamed then fried for 5 minutes each it will not be cooked, and at 15 minutes each it will be overcooked and burned on the surface. In either case, we won’t get any valid sensory ratings, because the products will be far beyond the ranges of acceptability.”

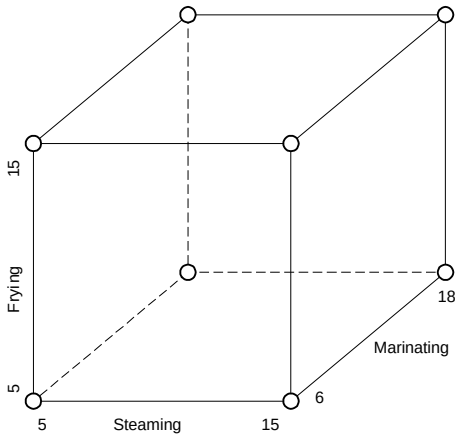
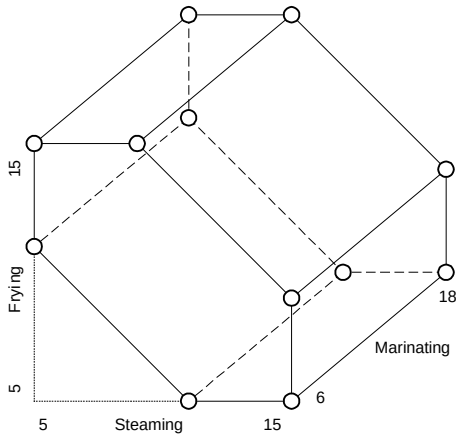
After some discussion, the process engineer and the statistician agree that an additional condition should be included:

“In order for the meat to be suitably cooked, the sum of the two cooking times should remain between 16 and 24 minutes for all experiments”.

This type of restriction is called a **multi-linear constraint**. In the current case, it can be written in a mathematical form requiring two equations, as follows:

Steam + Fry >= 16 and Steam + Fry <= 24

The impact of these constraints on the shape of the experimental region is shown in the two figures hereafter:

The cooked meat experimental region - no constraint	The cooked meat experimental region - multi-linear constraints
	

The constrained experimental region is no longer a cube! As a consequence, it is impossible to build a full factorial design in order to explore that region.

The design that best spans the new region is given in the table hereafter.

The cooked meat constrained design

Sample	Mar. Time	Steam. Time	Fry. Time
1	6	5	11
2	6	5	15

3	6	9	15
4	6	11	5
5	6	15	5
6	6	15	9
7	18	5	11
8	18	5	15
9	18	9	15
10	18	11	5
11	18	15	5
12	18	15	9

As you can see, it contains all "corners" of the experimental region, in the same way as the full factorial design does when the experimental region has the shape of a cube.

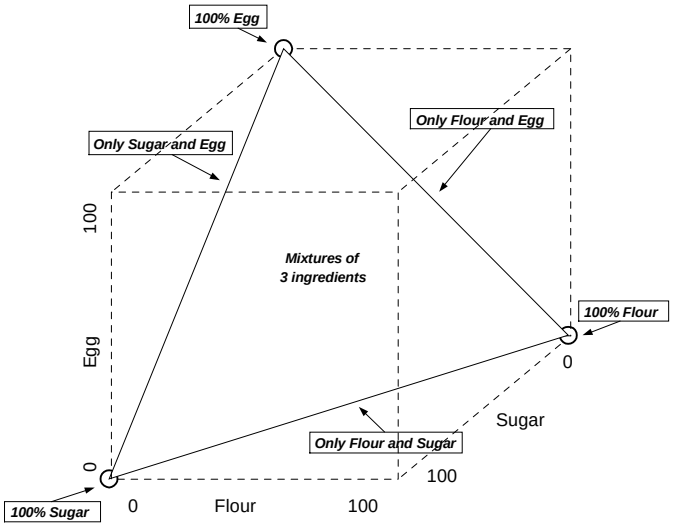
Depending on the number and complexity of multi-linear constraints to be taken into account, the shape of the experimental region can be more or less complex. In the worst cases, it may be almost impossible to imagine! Therefore, building a design to screen or optimize variables linked by multi-linear constraints requires special methods. Chapter “Alternative Solutions” below will briefly introduce two ways to build constrained designs.

A Special Case: Mixture Situations

A colleague of our process engineer, working in the Product Development department, has a different problem to solve: optimize a pancake mix. The mix consists of the following ingredients: wheat flour, sugar and egg powder. It will be sold in retail units of 100 g, to be mixed with milk for reconstitution of pancake dough.

The product developer has learnt about experimental design, and tries to set up an adequate design to study the properties of the pancake dough as a function of the amounts of flour, sugar and egg in the mix. She starts by plotting the region that encompasses all possible combinations of those three ingredients, and soon discovers that it has quite a peculiar shape:

The pancake mix experimental region



The reason, as you will have guessed, is that the mixture always has to add up to a total of 100 g. This is a special case of multi-linear constraint, which can be written with a single equation:

Flour + Sugar + Egg = 100

This is called the **mixture constraint**: the sum of all mixture components is 100% of the total amount of product.

The practical consequence, as you will also have noticed, is that the mixture region defined by three ingredients is not a three-dimensional region! It is contained in a two-dimensional surface called a **simplex**.

Therefore, mixture situations require specific designs. Their principles will be introduced in the next chapter.

Alternative Solutions

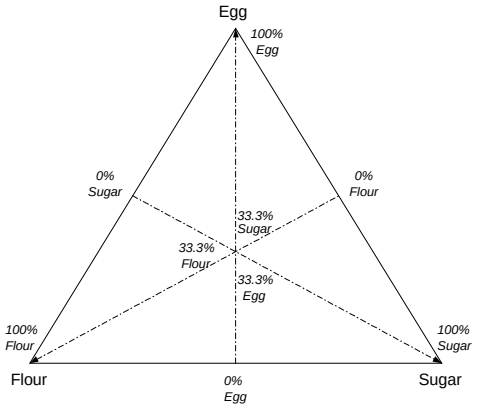
There are several ways to deal with constrained experimental regions. We are going to focus on two well known, proven methods:

- **Classical mixture designs** take advantage of the regular simplex shape that can be obtained under favorable conditions.
- In all other cases, a design can be computed algorithmically by applying the **D-optimal principle**.

Designs based on a simplex

Let us continue with the pancake mix example. We will have a look at the pancake mix simplex from a very special point of view. Since the region defined by the three mixture components is a two-dimensional surface, why not forget about the original three dimensions and focus only on this triangular surface?

The pancake mix simplex

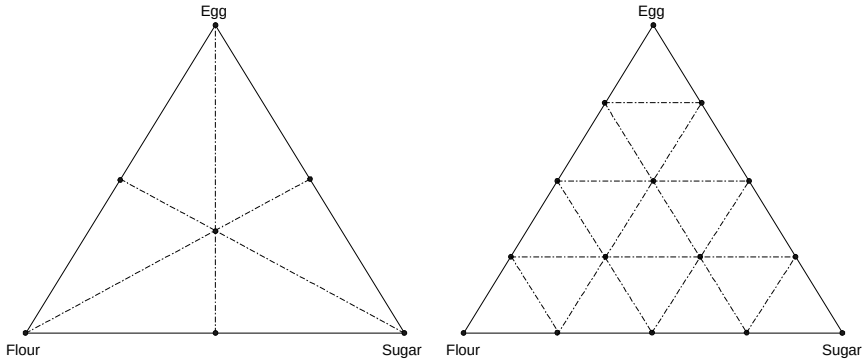


This simplex contains all possible combinations of the three ingredients flour, sugar and egg. As you can see, it is completely symmetrical. You could substitute egg for flour, sugar for egg and flour for sugar in the figure, and still get exactly the same shape.

Classical mixture designs take advantage of this symmetry. They include a varying number of experimental points, depending on the purposes of the investigation. But whatever this purpose and whatever the total number of experiments, these points are always symmetrically distributed, so that all mixture variables play equally important roles. These designs thus ensure that the effects of all investigated mixture variables will be studied with the same precision. This property is equivalent to the properties of factorial, central composite or Box-Behnken designs for non-constrained situations.

The figure hereafter shows two examples of classical mixture designs.

Two classical designs for 3 mixture components



The first design is very simple. It contains three corner samples (pure mixture components), three edge centers (binary mixtures) and only one mixture of all three ingredients, the centroid.

The second one contains more points, spanning the mixture region regularly in a triangular lattice pattern. It contains all possible combinations (within the mixture constraint) of five levels of each ingredient. It is similar to a 5-level full factorial design - except that many combinations, such as "25%,25%,25%" or "50%,75%,100%", are excluded because they are outside the simplex.

Read more about classical mixture designs in Chapter “Designs for Simple Mixture Situations” p.30.

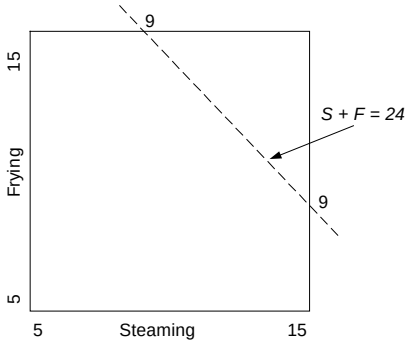
D-optimal designs

Let us now consider the meat example again (see Chapter “Constraints Between the Levels of Several Design Variables” p.25), and simplify it by focusing on Steaming time and Frying time, and taking into account only one constraint:

Steaming time + Frying time <= 24.

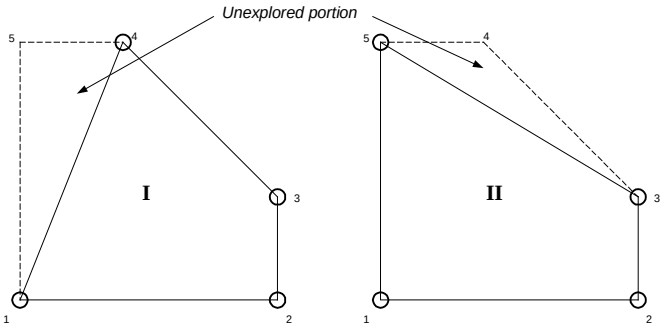
The figure hereafter shows the impact of the constraint on the variations of the two design variables.

The constraint cuts off one corner of the "cube"



If we try to build a design with only 4 experiments, as in the full factorial design, we will automatically end up with an imperfect solution that leaves a portion of the experimental region unexplored. This is illustrated in the next figure.

Designs with 4 points leave out a portion of the experimental region



On the figure, design II is better than design I, because the left out area is smaller. A design using points (1,3,4,5) would be equivalent to (I), and a design using points (1,2,4,5) would be equivalent to (II). The worst solution would be a design with points (2,3,4,5): it would leave out the whole corner defined by points 1,2 and 5.

Thus it becomes obvious that, if we want to explore the whole experimental region, we need more than 4 points. Actually, in the above example, the five points (1,2,3,4,5) are necessary. These five crucial points are the **extreme vertices** of the constrained experimental region. They have the following property: if you were to wrap a sheet of paper around those points, the shape of the experimental region would appear, revealed by your wrapping.

When the number of variables increases and more constraints are introduced, it is not always possible to include all extreme vertices into the design. Then you need a decision rule to select the best possible subset of points to include in your design. There are many possible rules; one of them is based on the so-called **D-optimal principle**, which consists in enclosing **maximum volume** into the selected points. In other words, you know that a wrapping of the selected points will not exactly re-constitute the experimental region you are interested in, but you want to leave out the smallest possible portion.

Read more about D-optimal designs and their various applications in Chapter Introduction to the D-Optimal Principle p.35.

Designs for Simple Mixture Situations

This chapter addresses the classical mixture case, where at least three ingredients are combined to form a blend, and three additional conditions are fulfilled:

1. The total amount of the blend is fixed (e.g. 100%);
2. There are no other constraints linking the proportions of two or more of the ingredients;
3. The ranges of variation of the proportions of the mixture ingredients are such that the experimental region has the regular shape of a simplex (see “Chapter Is the Mixture Region a Simplex?” p.49).

These conditions will be clarified and illustrated by an example. Then three possible applications will be considered, and the corresponding designs will be presented.

An Example of Mixture Design

This example, taken from John A. Cornell’s reference book “Experiments With Mixtures”, illustrates the basic principles and specific features of mixture designs.

A fruit punch is to be prepared by blending three types of fruit juice: watermelon, pineapple and orange. The purpose of the manufacturer is to use their large supplies of watermelons by introducing watermelon juice, of little value by itself, into a blend of fruit juices. Therefore, the fruit punch has to contain a substantial amount of watermelon - at least 30% of the total. Pineapple and orange have been selected as the other components of the mixture, since juices from these fruits are easy to get and inexpensive.

The manufacturer decides to use experimental design to find out which combination of those three ingredients maximizes consumer acceptance of the taste of the punch. The ranges of variation selected for the experiment are as follows:

Ranges of variation for the fruit punch design

Ingredient	Low	High	Centroid
Watermelon	30%	100%	54%
Pineapple	0%	70%	23%
Orange	0%	70%	23%

You can see at once that the resulting experimental design will have a number of features that make it very different from a factorial or central composite design.

Firstly, the ranges of variation of the three variables are not independent. Since Watermelon has a low level of 30%, the high level of Pineapple cannot be higher than 100 - 30 = 70%. The same holds for Orange.

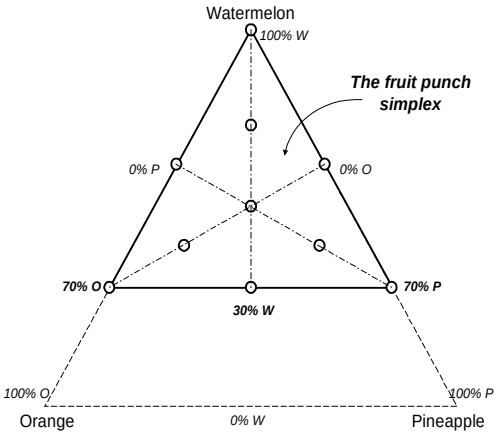
The second striking feature concerns the levels of the three variables for the point called “centroid”: these levels are not half-way between “low” and “high”, they are closer to the low level. The reason is, once again, that the blend has to add up to a total of 100%.

Since the levels of the various concentrations of ingredients to be investigated cannot vary independently from each other, these variables cannot be handled in the same way as the design variables encountered in a factorial or central composite design. To mark this difference, we will refer to those variables as **mixture components** (or mixture variables).

Whenever the low and high levels of the mixture components are such that the mixture region is a simplex (as shown in Chapter “A Special Case: Mixture Situations” p.27), classical mixture designs can be built. Read more about the necessary conditions in Chapter “Is the Mixture Region a Simplex?” p.49.

These designs have a fixed shape, depending only on the number of mixture components and on the objective of your investigation. For instance, we can build a design for the optimization of the concentrations of Watermelon, Pineapple and Orange juice in Cornell's fruit punch, as shown in the figure below.

Design for the optimization of the fruit punch composition



The next chapters will introduce the three types of mixture designs that are most suitable for three different objectives:

1. ***Screening*** of the effects of several mixture components;

2. ***Optimization*** of the concentrations of several mixture components;
3. ***Even coverage*** of an experimental region.

Screening Designs for Mixtures

In a screening situation, you are mostly interested in studying the main effects of each of your mixture components.

What is the best way to build a mixture design for screening purposes? To answer this question, let us go back to the concept of main effect.

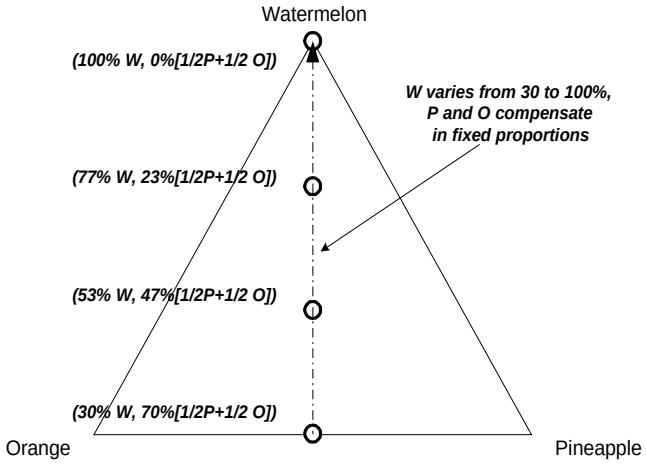
The main effect of an input variable on a response is the change occurring in the response values when the input variable varies from Low to High, all experimental conditions being otherwise comparable.

In a factorial design, the levels of the design variables are combined in a balanced way, so that you can follow what happens to the response value when a particular design variable goes from Low to High. It is mathematically possible to compute the main effect of that design variable, because its Low and High levels have been combined with the same levels of all the other design variables.

In a mixture situation, this is no longer possible. Look at the Fruit Punch image above: while 30% Watermelon can be combined with (70% P, 0% O) and (0% P, 70% O), 100% Watermelon can only be combined with (0% P, 0% O)!

To find a way out of this dead end, we have to transpose the concept of "otherwise comparable conditions" to the constrained mixture situation. To follow what happens when Watermelon varies from 30% to 100%, let us compensate for this variation in such a way that the mixture still adds up to 100%, without disturbing the balance of the other mixture components. This is achieved by moving along an axis where the proportions of the other mixture components remain constant, as shown in the figure below.

Studying variations in the proportion of Watermelon

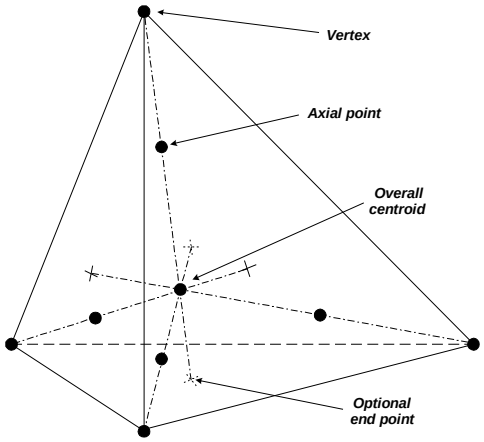


The most "representative" axis to move along is the one where the other mixture components have equal proportions. For instance, in the above figure, Pineapple and Orange each use up one half of the remaining volume once Watermelon has been determined.

Mixture designs based upon the axes of the simplex are called **axial designs**. They are the best suited for screening purposes because they manage to capture the main effect of each mixture component in a simple and economical way.

A more general type of axial design is represented, for four variables, in the next figure. As you can see, most of the points are located inside the simplex: they are mixtures of all four components. Only the four corners, or vertices (containing the maximum concentration of an individual component) are located on the surface of the experimental region.

A 4-component axial design

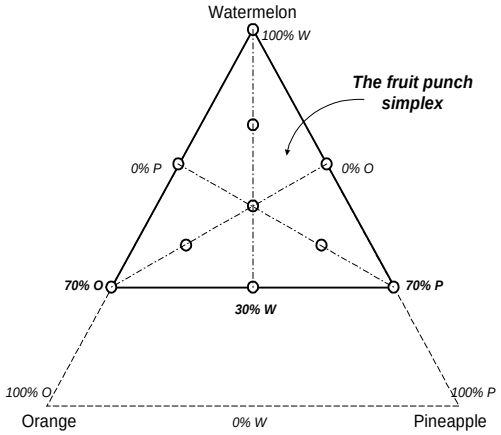


Each **axial point** is placed halfway between the overall centroid of the simplex (25%,25%,25%,25%) and a specific vertex. Thus the path leading from the centroid ("neutral" situation) to a vertex (extreme situation with respect to one specific component) is well described with the help of the axial point.

In addition, **end points** can be included; they are located on the surface of the simplex, opposite to a vertex (the are marked by crosses on the figure). They contain the minimum concentration of a specific component. When end points are included in an axial design, the whole path leading from minimum to maximum concentration is studied.

The Fruit Punch Mixture Region

Design for the optimization of the fruit punch composition



Optimization Designs for Mixtures

If you wish to optimize the concentrations of several mixture components, you need a design that enables you to predict with a high accuracy what happens for any mixture - whether it involves all components or only a subset.

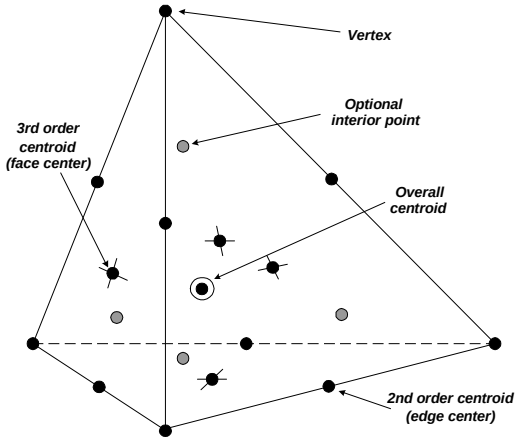
It is a well-known fact that peculiar behaviors often happen when a concentration drops down to zero. For instance, to prepare the base for a Dijon mayonnaise, you need to blend Dijon mustard, egg and vegetable oil. Have you ever tried - or been forced by circumstances - to remove the egg from the recipe? If you do, you will get a dressing with a different appearance and texture. This illustrates the importance of interactions (e.g. between egg and oil) in mixture applications.

Thus, an optimization design for mixtures will include a large number of blends of only two, three, or more generally a subset of the components you want to study. The most regular design including those sub-blends is called **simplex-centroid design**. It is based on the **centroids** of the simplex: balanced blends of a subset of the mixture components of interest. For instance, to optimize the concentrations of three ingredients, each of them varying between 0 and 100%, the simplex-centroid design will consist of:

- The 3 vertices: (100,0,0), (0,100,0) and (0,0,100);
- The 3 edge centers (or centroids of the 2-dimensional sub-simplexes defining binary mixtures): (50,50,0), (50,0,50) and (0,50,50);
- The overall centroid: (33,33,33).

A more general type of simplex-centroid design is represented, for 4 variables, in the figure below.

A 4-component simplex-centroid design



If all mixture components vary from 0 to 100%, the blends forming the simplex-centroid design are as follows:

1- The vertices are pure components;

2- The second order centroids (edge centers) are binary mixtures with equal proportions of the selected two components;

3- The third order centroids (face centers) are ternary mixtures with equal proportions of the selected three components;

…..

N- The overall centroid is a mixture where all N components have equal proportions.

In addition, **interior points** can be included in the design. They improve the precision of the results by "anchoring" the design with additional complete mixtures. The most regular design is obtained by adding

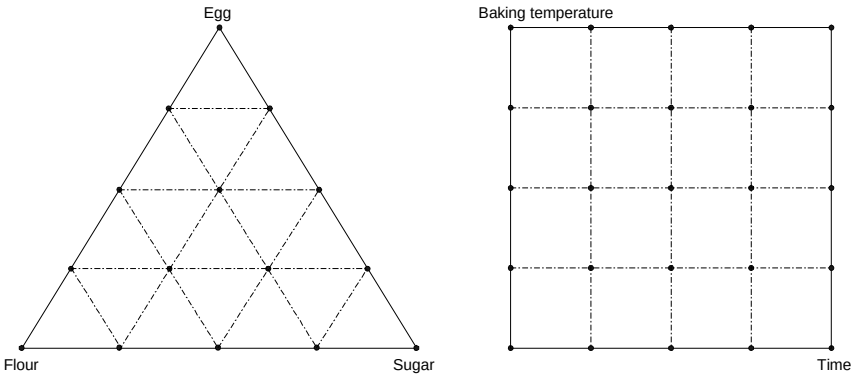
interior points located halfway between the overall centroid and each vertex. They have the same composition as the axial points in an axial design.

Designs that Cover a Mixture Region Evenly

Sometimes you may not be specifically interested in a screening or optimization design. In fact, you may not even know whether you are ready for a screening! For example, you just want to investigate what would happen if you mixed three ingredients that you have never tried to mix before.

This is one of the cases when your main purpose is to cover the mixture region as evenly and regularly as possible. Designs that address that purpose are called **simplex-lattice** designs. They consist of a network of points located at regular intervals between the vertices of the simplex. Depending on how thoroughly you want to investigate the mixture region, the network will be more or less dense, including a varying number of intermediate levels of the mixture components. As such, it is quite similar to an N-level full factorial design. The figure below illustrates this similarity.

A 4th degree simplex-lattice design is similar to a 5-level full factorial



In the same way as a full factorial design, depending on the number of levels, can be used for screening, optimization, or other purposes, simplex-lattice designs have a wide variety of applications, depending on their **degree** (number of intervals between points along the edge of the simplex). Here are a few:

- ***Feasibility study*** (degree 1 or 2): are the blends feasible at all?
- ***Optimization***: with a lattice of degree 3 or more, there are enough points to fit a precise response surface model.
- Search for a ***special behavior*** or property which only occurs in an unknown, limited sub-region of the simplex.
- ***Calibration***: prepare a set of blends on which several types of properties will be measured, in order to fit a regression model to these properties. For instance, you may wish to relate the texture of a product, as assessed by a sensory panel, to the parameters measured by a texture analyzer. If you know that texture is likely to vary as a function of the composition of the blend, a simplex-lattice design is probably the best way to generate a representative, balanced calibration data set.

Introduction to the D-Optimal Principle

If you are familiar with factorial designs, you probably know that their most interesting feature is that they allow you to study all effects **independently from each other**. This property, called **orthogonality**, is vital for relating variations of the responses to variations in the design variables. It is what allows you to draw conclusions about cause and effect relationships. It has another advantage, namely **minimizing the error** in the estimation of the effects.

Constrained Designs Are Not Orthogonal

As soon as Multi-Linear Constraints are introduced among the design variables, it is no longer possible to build an orthogonal design. This can be grasped intuitively if you understand that orthogonality is equivalent to the fact that all design variables are varied independently from each other. As soon as the variations in one of the design variables are linked to those of another design variable, orthogonality cannot be achieved.

In order to minimize the negative consequences of a deviation from the ideal orthogonal case, you need a measure of the "lack of orthogonality" of a design. This measure is provided by the **condition number**, defined as follows:

Cond# = square root (largest eigenvalue / smallest eigenvalue)

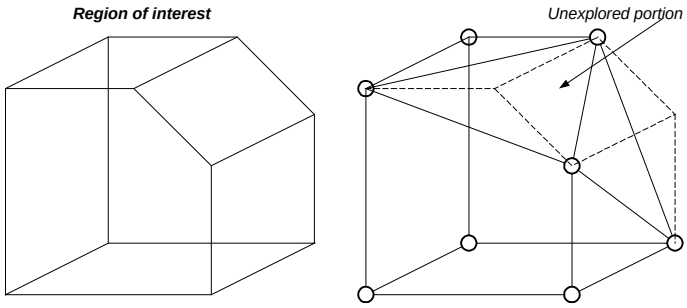
which is linked to the elongation or degree of "non-sphericity" of the region actually explored by the design. The smaller the condition number, the more spherical the region, and the closer you are to an orthogonal design.

Small Condition Number Means Large Enclosed Volume

Another important property of an experimental design is its ability to explore the whole region of possible combinations of the levels of the design variables. It can be shown that, once the shape of the experimental region has been determined by the constraints, the design with the smallest condition number is the one that encloses **maximal volume**.

In the ideal case, if all extreme vertices are included into the design, it has the smallest attainable condition number. If that solution is too expensive, however, you will have to make a selection of a smaller number of points. The automatic consequence is that the condition number will increase and the enclosed volume will decrease. This is illustrated by the next figure.

With only 8 points, the enclosed volume is not optimal



How a D-Optimal Design Is Built

First, the purpose of the design has to be expressed in the form of a mathematical **model**. The model does not have the same shape for a screening design as for an optimization design.

Once the model has been fixed, the **condition number** of the "experimental matrix", which contains one column per effect in the model, and one row per experimental point, can be computed.

The **D-optimal algorithm** will then consist in:

1. Deciding how many points the design should include. Read more about that in chapter “How Many Experiments Are Necessary?” p.51.
2. Generating a set of **candidate points**, among which the points of the design will be selected. The nature of the relevant candidate points depends on the shape of the model. Read the next chapters for more details.

3. Selecting a subset with the desired number of points more or less randomly, and computing the condition number of the resulting experimental matrix.
4. Exchanging one of the selected points with a left over point and comparing the new condition number to the previous one. If it is lower, the new point replaces the old one; else another left over point is tried. This process can be re-iterated a large number of times.

When the exchange of points does not give any further improvements, the algorithm stops and the subset of candidate points giving the lowest condition number is selected.

How Good Is My Design?

The excellence of a D-optimal design is expressed by its condition number, which, as we have seen previously, depends on the shape of the model as well as on the selected points.

In the simplest case of a linear model, an orthogonal design like a full factorial would have a condition number of 1. It follows that the condition number of a D-optimal design will always be larger than 1. A D-optimal design with a linear model is acceptable up to a cond# around 10.

If the model gets more complex, it becomes more and more difficult to control the increase in the condition number. For practical purposes, one can say that a design including interaction and/or square effects is usable up to a cond# around 50.

If you end up with a cond# much larger than 50 no matter how many points you include in the design, it probably means that your experimental region is too constrained. In such a case, it is recommended that you re-examine all of the design variables and constraints with a critical eye. You need to search for ways to simplify your problem (see Chapter “Advanced Topics for Constrained Situations” p.49), otherwise you run the risk of starting an expensive series of experiments which will not give you any useful information at all.

D-Optimal Designs Without Mixture Variables

D-optimal designs for situations that do not involve a blend of constituents with a fixed total will be referred to as "non-mixture" D-optimal designs. To differentiate them from mixture components, we will call the design variables involved in non-mixture designs **process variables**.

A non-mixture D-optimal design is the solution to your experimental design problem every time you want to investigate the effects of several process variables linked by one or more Multi-Linear Constraints. It is built according to the D-optimal principle described in the previous chapter.

D-Optimal Designs for Screening Stages

If your purpose if to focus on the main effects of your design variables, and optionally to describe some or all of the interactions among them, you will need a ***linear*** model, optionally with ***interaction effects***.

The set of candidate points for the generation of the D-optimal design will then consist mostly of the extreme vertices of the constrained experimental region. If the number of variables is small enough, edge centers and higher order centroids can also be included.

In addition, center samples are automatically included in the design (whenever they apply); they are not submitted to the D-optimal selection procedure.

D-Optimal Designs for Optimization Purposes

When you want to investigate the effects of your design variables with enough precision to describe a response surface accurately, you need a ***quadratic model***. This model requires intermediate points (situated somewhere between the extreme vertices) so that the square effects can be computed.

The set of candidate points for a D-optimal optimization design will thus include:

- all extreme vertices;
- all edge centers;
- all face centers and constraint plane centroids.

To imagine the result in three dimensions, you can picture yourself a combination of a Box-Behnken design (which includes all edge centers) and a Cubic Centered Faces design (with all corners and all face centers). The main difference is that the constrained region is not a cube, but a more complex polyhedron.

The D-optimal procedure will then select a suitable subset from these candidate points, and several replicates of the overall center will also be included.

D-Optimal Designs With Mixture Variables

The D-optimal principle can solve mixture problems in two situations:

1. The mixture region is not a simplex.
2. Mixture variables have to be combined with process variables.

Pure Mixture Experiments

When the mixture region is not a simplex (see Is the Mixture Region a Simplex?), a D-optimal design can be generated in a way similar to the process cases described in the previous chapter.

Here again, the set of candidate points depends on the shape of the model. You may lookup Chapter “Relevant Regression Models” in the section on analyzing results from designed experiments for more details on mixture models.

The overall centroid is always included in the design, and is not subject to the D-optimal selection procedure.

Note: Classical mixture designs have much better properties than D-optimal designs. Remember this before establishing additional constraints on your mixture components!

Chapter “How To Select Reasonable Constraints” p.50 tells you more about how to avoid unnecessary constraints.

How To Combine Mixture and Process Variables

Sometimes the product properties you are interested in depend on the combination of a mixture recipe with specific process settings. In such cases, it is useful to investigate mixture and process variables together.

The Unscrambler offers three different ways to build a design combining mixture and process variables. They are described below.

The mixture region is a simplex

When your mixture region is a simplex, you may combine a classical mixture design, as described in Chapter Designs for Simple Mixture Situations, with the levels of your process variables, in two different ways.

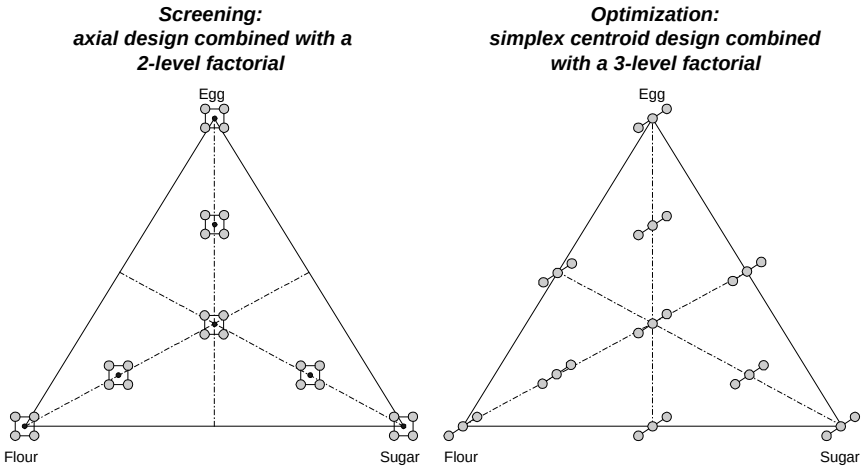
The first solution is useful when several process variables are included in the design. It applies the D-optimal algorithm to select a subset of the candidate points, which are generated by combining the complete mixture design with a full factorial in the process variables.

Note: The D-optimal algorithm will usually select only the extreme vertices of the mixture region. Be aware that the resulting design may not always be relevant!

The D-optimal solution is acceptable if you are in a screening situation (with a large number of variables to study) and the mixture components have a lower limit. If the latter condition is not fulfilled, the design will include only pure components, which is probably not what you had in mind!

The alternative is to use the whole set of candidate points. In such a design, each mixture is combined with all levels of the process variables. The figure below illustrates two such situations.

Two full factorial combinations of process variables with complete mixture designs



This solution is recommended (if the number of factorial combinations is reasonable) whenever it is important to explore the mixture region precisely.

The mixture region is not a simplex

If your mixture region is not a simplex, you have no choice: the design has to be computed by a D-optimal algorithm. The candidate points consist of combinations of the extreme vertices (and optionally lower-order centroids) with all levels of the process variables. From these candidate points, the algorithm will select a subset of the desired size.

Note: When the mixture region is not a simplex, only continuous process variables are allowed.

Various Types of Samples in Experimental Design

This section presents an overview of the various types of samples to be found in experimental design and their properties.

Cube Samples

Cube samples can be found in factorial designs and their extensions.

They are a combination of **high** and **low** levels of the design variables, in experimental plans based on two levels of each variable.

This also applies to Central Composite designs (they contain the full factorial cube).

More generally, all combinations of levels of the design variables in N-level full factorials, as well as in Simplex lattice designs, are also called cube samples.

In Box-Behnken designs, all samples that are a combination of high or low levels of some design variables, and center level of others, are also referred to as cube samples.

Center Samples

Center samples are samples for which each design variable is set at its mid-level. They are located at the exact center of the experimental region.

Center Samples in Screening Designs

In screening designs, center samples are used for ***curvature checking***: Since the underlying model in such a design assumes that all main effects are linear, it is useful to have at least one design point with an intermediate level for all factors. Thus, when all experiments have been performed, you can check whether the intermediate value of the response fits with the global linear pattern, or whether it is far from it (curvature). In the case of high curvature, you will have to build a new design that accepts a quadratic model.

In screening designs, center samples are optional; however, we recommend that you include at least two if possible.

See section Replicates p.43 for details about the use of replicated center samples.

Center Samples in Optimization Designs

Optimization designs automatically include at least one center sample, which is necessary as a kind of anchor point to the quadratic model. Furthermore, you are strongly recommended to include more than one. The default number of center samples for Central Composite and Box-Behnken designs is computed so as to achieve ***uniform precision*** all over the experimental region.

Sample Types in Central Composite Designs

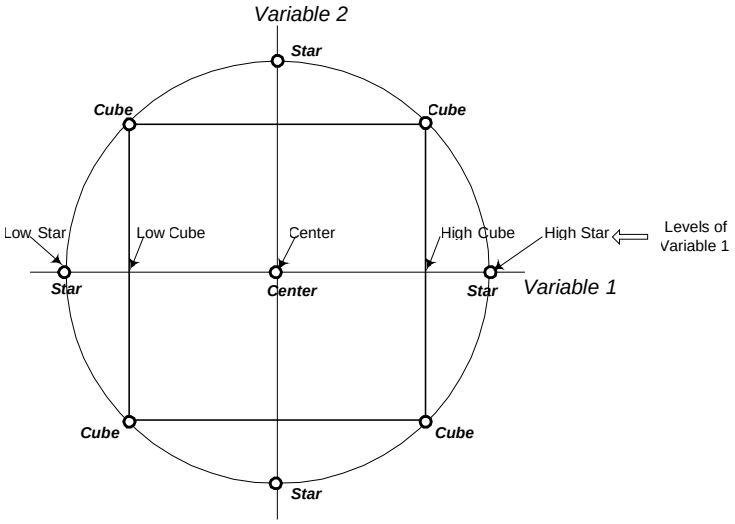
Central Composite designs include the following types of samples:

- Cube samples (see Cube Samples);
- Center samples (see Center Samples in Optimization Designs);
- Star samples.

Star Samples

Star samples are samples with ***mid-values*** for all design variables except one, for which the value is ***extreme***. They provide the necessary intermediate levels that will allow a quadratic model to be fitted to the data.

Star samples in a Central Composite design with two design variables



Star samples can be centers of cube faces, or they can lie outside the cube, at a given distance (larger than 1) from the center of the cube.

By default, their distance to the center is the same as the distance from the cube samples to the center, i.e. here:

$$\sqrt{2} \cdot \frac{\left(High\ Cube - Low\ Cube\right)}{2}$$

Distance To Center

The properties of the Central Composite design will vary according to the distance between the star samples and the center samples. This distance is measured in normalized units, i.e. assuming that the low cube level of each variable is -1 and the high cube level +1.

Three cases can be considered:

1. The ***default*** star distance to center ensures that ***all design samples are located on the surface of a sphere***. In other words, the star samples are as far away from the center as the cube samples are. As a consequence, all design samples have exactly the same leverage. The design is said to be "rotatable";
7. The star distance to center can be ***tuned down to 1***. In that case, the star samples will be located at the ***centers of the faces*** of the cube. This ensures that a Central Composite design can be built even if levels lower than "low cube" or higher than "high cube" are impossible. However, the design is ***no longer rotatable***;
8. Any ***intermediate value*** for the star distance to center is also possible. The design will ***not*** be ***rotatable***.

Sample Types in Mixture Designs

Here is an overview of the various sample types available in each type of classical mixture design:

- Axial design: vertex samples, axial points, optional end points, overall centroid;
- Simplex-centroid design: vertex samples, centroids of various orders, optional interior points, overall centroid ;
- Simplex-lattice designs: cube samples (see Cube Samples), overall centroid.

Each type is described hereafter.

Axial Point

In an axial design, an ***axial point*** is positioned on the axis of one of the mixture variables, and must be above the overall centroid, opposite the end point.

Centroid Point

A ***centroid point*** is calculated as the mean of the extreme vertices on a given surface. Edge centers, face centers and overall centroid are all examples of centroid points.

The number of mixture components involved in the centroid is called the centroid ***order***. For instance, in a 4-component mixture, the overall centroid is the fourth order centroid.

Edge Center

The ***edge centers*** are positioned in the center of the edges of the simplex. They are also referred to as ***second order centroids***.

End Point

In an axial or a simplex-centroid design, an ***end point*** is positioned at the bottom of the axis of one of the mixture variables, and is thus on the opposite side to the axial point.

Face Center

The ***face centers*** are positioned in the center of the faces of the simplex. They are also referred to as ***third order centroids***.

Interior Point

An ***interior point*** is not located on the surface, but inside the experimental region. For example, an axial point is a particular kind of interior point.

Overall Centroid

The ***overall centroid*** is calculated as the mean of all extreme vertices. It is the mixture equivalent of a center sample.

Vertex Sample

A ***vertex*** is a point where two lines meet to form an angle. Vertex samples are the “corners” of D-optimal or mixture designs.

Sample Types in D-Optimal Designs

D-optimal designs may contain the following types of samples:

- ***vertex samples***, also called ***extreme vertices*** (see the description of a Vertex Sample above);
- ***centroid points*** (see Centroid Point, Edge Center and Face Center);
- ***overall centroid*** (see Overall Centroid).

Reference Samples

Reference samples are experiments which do not belong to a standard design, but which you choose to include for various purposes.

Here are a few classical cases where reference samples are often used:

- If you are trying to ***improve*** an existing product or process, you might use the ***current*** recipe or process settings as reference.
- If you are trying to ***copy*** an ***existing product***, for which you do not know the recipe, you might still include it as reference and measure your responses on that sample as well as on the others, in order to know how close you have come to that product.
- To ***check curvature*** in the case where some of the design variables are category variables, you can include one reference sample with ***center levels of all continuous variables*** for each level (or combination of levels) of the category variable(s).

Note: For reference samples, only response values can be taken automatically into account in the Analysis of Effects and Response Surface analyses. You may, however, enter the values of the design variables manually after converting to non-designed data table, then run a PLS analysis.

Replicates

Replicates are experiments performed several times. They should not be confused with repeated measurements, where the samples are only prepared once but the measurements are performed several times on each.

Why Include Replicates?

Replicates are included in a design in order to make estimation of the ***experimental error*** possible. This is doubly useful:

- It gives information about the ***average experimental error*** in itself;
- It enables you to compare response variation due to ***controlled causes*** (i.e. due to variation in the design variables) with ***uncontrolled*** response variation. If the "explainable" variation in a response is no larger than its random variation, the variations of this response cannot be related to the investigated design variables.

How to Include Replicates

The usual strategy is to specify several replicates of the ***center sample***. This has the advantage of both being rather economical, and providing you with an estimation of the experimental error in "average" conditions.

When no center sample can be defined (because the design includes category variables or variables with more than two levels), you may specify replicates for one or several ***reference samples*** instead.

But if you know that there is a lot of uncontrolled or unexplained variability in your experiments, it might be wise to ***replicate the whole design***, i.e. to perform all experiments twice.

Sample Order in a Design

The purpose of experimental design usually is to find out how variations in design variables influence response variations. However we know that, no matter how well we strive to control the conditions of our experiments, random variations still occur. The next sections describe what can be done to limit the effect of random variations on the interpretation of the final results.

Randomization

Randomization means that the experiments are performed in random order, as opposed to the standard order which is sorted according to the levels of the design variables.

Why Is Randomization Useful?

Very often, the experimental conditions are likely to vary somewhat in time along the course of the investigation, such as when temperature and humidity vary according to external meteorological conditions, or when the experiments are carried out by a new employee who is better trained at the end of the investigation than at the beginning. It is crucial not to risk confusing the effect of a change over time with the effect of one of the investigated variables. To avoid such misinterpretation, the order in which the experimental runs are to be performed is usually randomized.

Incomplete Randomization

There may be circumstances which prevent you from using full randomization. For instance, one of the design variables may be a parameter that is particularly difficult to tune, so that the experiments will be performed much more efficiently if you only need to tune that parameter a few times. Another case for incomplete randomization is blocking (see Chapter Blocking hereafter).

The Unscrambler enables you to leave some variables out of the randomization. As a result, the experimental runs will be sorted according to the non-randomized variable(s). This will generate groups of samples with a constant value for those variables. Inside each such group, the samples will be randomized according to the remaining variables.

Blocking

In cases where you suspect experimental conditions to vary from time to time or from place to place, and when only some of the experiments can be performed under constant conditions, you may consider to use ***blocking*** of your set of experiments instead of free randomization. This means that you incorporate an extra design variable for the blocks. Experimental runs must then be randomized within each block.

Typical examples of blocking factors are:

- ***Day*** (if several experimental runs can be performed the same day);
- ***Operator*** or ***machine*** or ***instrument*** (when several of them must be used in parallel to save time);
- ***Batches*** (or shipments) of raw material (in case one batch is insufficient for all runs).

Blocking is not handled automatically in The Unscrambler, but it can be done manually using one or several additional design variables. Those variables should be left out of the randomization.

Extending a Design

Once you have performed a series of designed experiments, analyzed their results, and drawn a conclusion from them, two situations can occur:

1. The experiments have provided you with all the information you needed, which means that your project is ***completed***.
9. The experiments have given you valuable information which you can use to ***build a new series*** of experiments that will lead you closer to your objective.

In the latter case, the new series of experiments can sometimes be designed as a complement to, or an extension of, the previous design. This lets you minimize the number of new experimental runs, and the whole set of results from the two series of runs can be analyzed together.

Why Extend A Design?

In principle, you should make use of the extension feature whenever possible, because it enables you to go one step further in your investigations with a minimum of additional experimental runs, since it takes into account the already performed experiments.

Extending an existing design is also a nice way to build a new, similar design that can be analyzed together with the original one. For instance, if you have investigated a reaction using a specific type of catalyst, you might want to investigate another type of catalyst in the same conditions as the first one in order to compare their performances. This can be achieved by adding a new design variable, namely type of catalyst, to the existing design.

You can also use extensions as a basis for an efficient sequential experimental strategy. That strategy consists in breaking your initial problem into a series of smaller, intermediate problems and invest into a small number of experiments to achieve each of the intermediate objectives. Thus, if something goes wrong at one stage, the losses are cut, and if all goes well, you will end up solving the initial problem at a lower cost than if you had started off with a huge design.

Which Designs Can Be Extended?

Full and fractional factorial designs, central composite designs, D-optimal designs and mixture designs can be extended in various manners.

The tables hereafter list the possible types of extensions and the designs they apply to:

Types of extensions for orthogonal designs

Type of extension	Fractional Factorial	Full Factorial	CCD
Add levels	No	Yes	No
Add a design variable	Yes	Yes	No
Delete a design variable	Yes	Yes	No
Add more replicates	Yes	Yes	Yes
Add more center samples	Yes(*)	Yes(*)	Yes
Add more reference samples	Yes	Yes	Yes
Extend to higher resolution	Yes	-	-
Extend to full factorial	Yes	-	-
Extend to central composite	Yes(*)	Yes(*)	-

(*) Applies to 2-level continuous variables only.

Types of extensions for D-optimal and Mixture designs

Type of extension	D-opt Non mixture	Mixture with Process	Lattice (no Process)	Centroid (no Process)	Axial (no Process)
Add levels to Process Variables	No	Yes(**)	-	-	-
Add more replicates	Yes	Yes	Yes	Yes	Yes
Add more center samples	Yes	Yes	Yes	Yes	Yes
Add more reference samples	Yes	Yes	Yes	Yes	Yes
Increase lattice degree	-	No	Yes	-	-
Extend to centroid	-	No	Yes	-	Yes
Add interior points	-	No	-	Yes	-
Add end points	-	No	-	-	Yes

(**) Only if experimental region is a simplex.

In addition, all designs which are not listed in the above tables can be extended by adding more center and reference samples or replicates.

When and How To Extend A Design

Let us now go briefly through the most common extension cases:

- ***Add levels***: Used whenever you are interested in investigating more levels of already included design variables, especially for category variables.
- ***Add a design variable***: Used whenever a parameter that has been kept constant is suspected to have a potential influence on the responses, as well as when you wish to duplicate an existing design in order to apply it to new conditions that differ by the values of one specific variable (continuous or category), and analyze the results together. For instance, you have just investigated a chemical reaction using a specific catalyst, and now wish to study another similar catalyst for the same reaction and compare its performances to the other one's. The simplest way to do this is to extend the first design by adding a new variable; type of catalyst.
- ***Delete a design variable***: If the analysis of effects has established one or a few of the variables in the original session to be clearly non-significant, you can increase the power of your conclusions by deleting this variable and reanalyzing the design. Deleting a design variable can also be a first step before extending a screening design into an optimization design. You should use this option with caution if the effect of the removed variable is close to significance. Also make sure that the variable you intend to remove does not participate in any significant interactions.
- ***Add more replicates***: If the first series of experiments shows that the experimental error is unexpectedly high, replicating all experiments once more might make your results clearer.
- ***Add more center samples***: If you wish to get a better estimation of the experimental error, adding a few center samples is a good and inexpensive solution.
- ***Add more reference samples***: Whenever new references are of interest, or if you wish to include more replicates of the existing reference samples in order to get a better estimation of the experimental error.
- ***Extend to higher resolution***: Use this option for fractional factorial designs where some of the effects you are interested in are confounded with each other. You can use this option whenever some of the confounded interactions are significant and you wish to find out exactly which ones. This is only possible if there is a higher resolution fractional factorial design. Otherwise, you can extend to full factorial instead.
- ***Extend to full factorial***: This applies to fractional factorial designs where some of the effects you are interested in are confounded with each other and no higher resolution fractional factorial designs are possible.

- ***Extend to central composite***: This option completes a full factorial design by adding star samples and (optionally) a few more center samples. Fractional factorial designs can also be completed this way, by adding the necessary cube samples as well. This should be used only when the number of design variables is small; an intermediate step may be to delete a few variables first.

Caution! Whichever kind of extension you use, remember that all the experimental conditions not represented in the design variables *must be the same* for the new experimental runs as for the previous runs.

Building an Efficient Experimental Strategy

How should you use experimental design in practice? Is it more efficient to build one global design that tries to achieve your main goal, or would it be better to break it down into a sequence of more modest objectives, each with its own design?

We strongly advise you, even if the initial number of design variables you wish to investigate is rather small, to use the latter, ***sequential approach***. This has at least four advantages:

1. Each step of the strategy consists of a design involving a reasonably small number of experiments. Thus, the mere size of each sub-project is ***more easily manageable***.
10. A smaller number of experiments also means that the underlying conditions more easily can be kept constant for the whole design, which will make the ***effects*** of the design variables appear ***more clearly***.
11. If something ***goes wrong*** at a given step, the damage is restricted to that particular step.
12. If all goes well, the ***global cost*** is usually ***smaller*** than with one huge design, and the final objective is achieved all the same.

Example of Experimental Strategy

Let us illustrate this with the following example. You wish to optimize a process that relies on 6 parameters: A, B, C, D, E, F. You do not know which of those parameters really matter, so you have to start from the screening stage.

The most ***straightforward approach*** would be to try an optimization at once, by building a CCD with 6 design variables. It is possible, but **costly** (at least 77 samples required) and **risky** (what happens if something goes wrong, like a wrong choice of ranges of variation? All experiments are lost).

Here is an ***alternative approach*** (note that the results mentioned hereafter only have illustrative value – in real life the number of significant results and their nature may be different):

1. First, you build a fractional factorial design 2^{6-2} (resolution IV), with 2 center samples, and you perform the corresponding 18 experiments.
2. After analyzing the results, it turns out (for example) that only variables A, B, C and E have significant main effects and/or interactions. But those interactions are confounded, so you need to extend the design in order to know which are really significant.
3. You extend the first design by deleting variables D and F and extending the remaining part (which is now a 2^{4-1}, resolution IV design) to a full factorial design with one more center sample. Additional cost: 9 experiments.
4. After analyzing the new design, the significant interactions which are not confounded only involve (for example) A, B and C. The effect of E is clear and goes in the same direction for all responses. But since your center samples show some curvature, you need to go to optimization stage for the remaining variables.
5. Thus, you keep variable E constant at its most interesting level, and after deleting that variable from the design you extend the remaining 2^3 full factorial to a CCD with 6 center samples. Additional cost: 9 experiments.

6. Analysis of the final results provides you (if all goes well) with a nice optimum. Final cost: 18+9+9=36 experiments, which is less than half of the initial estimate.

Advanced Topics for Unconstrained Situations

In the following section, you will find a few tips that might come in handy when you consider building a design or analyzing designed data.

How To Select Design Variables

Choosing which variables to investigate is the first step in designing experiments. That problem is best tackled during a ***brainstorming session*** in which all people involved in the project should participate, so as to make sure that no important aspect of the problem is forgotten.

- For a **first screening**, the most important rule is: ***Do not leave out a variable that may have an influence on the responses unless you know that you cannot control it in practice.*** It would be more costly to have to include one more variable at a later stage than to include one more in the first screening design.
- For a more extensive **screening**, variables that are known not to interact with other variables can be left out. If those variables have a negligible linear effect, you can choose whatever constant value you wish for them (e.g. the least expensive). If those variables have a significant linear effect, they should be fixed at the level most likely to give the desired effect on the response.
- The previous rule also applies to **optimization** designs, if you also know that the variables in question have no quadratic effect. If you suspect that a variable can have a non-linear effect, you should include it in the optimization stage.

How To Select Ranges of Variation

Once you have decided which variables to investigate, appropriate ***ranges of variation*** remain to be defined.

For **screening** designs, you are generally interested in covering the largest possible region. On the other hand, no information is available in the regions between the levels of the experimental factors unless you assume that the response behaves smoothly enough as a function of the design variables. Selecting the adequate levels is a trade-off between these two aspects.

Thus a ***rule of thumb*** can be applied: Make the range large enough to give effect and small enough to be realistic. If you suspect that two of the designed experiments will give extreme, opposite results, perform those first. If the two results are indeed different from each other, this means that you have generated enough variation. If they are too far apart, you have generated too much variation, and you should shrink the ranges a bit. If they are too close, try a center sample; you might just have a very strong curvature!

Since **optimization** designs are usually built after some kind of screening, you should already know roughly in what area the optimum lies. So unless you are building a CCD as an extension of a previous factorial design, you should try to select a ***smaller*** range of variation. This way a quadratic model will be more likely to approximate the true response surface correctly.

Model Validation for Designed Data Tables

In a screening design, if all possible interactions are present, each cube sample carries unique information. In such cases, if there are no replicates, the idea behind ***cross-validation*** is not valid, and usually the cross validation error will be very large.

Leverage correction is no better solution: For MLR-based methods, leverage correction is strictly equivalent to full cross validation, whereas it provides only rough estimates which cannot be trusted completely for projection methods, since leverage correction makes no actual predictions. An alternative validation method for such data is probability plotting of the principal component scores.

However, in other cases when there are several residual degrees of freedom in the cube and/or star samples, full cross validation can be used without trouble. This applies whenever the number of cube and/or star samples is much larger than the number of effects in the model.

The Importance of Having Measurements for All Design Samples

Analysis of effects and response surface modeling, which are specially tailored for orthogonally designed data sets, can only be run if response values are available for all the designed samples. The reason is that those methods need ***balanced data*** to be applicable. As a consequence, you should be especially careful to collect response values for all experiments. If you do not, for instance due to some instrument failure, it might be advisable to re-do the experiment later to collect the missing values.

If, for some reason, some response values simply cannot be measured, you will still be able to use the standard multivariate methods described in this manual: PCA on the responses, and PCR or PLS to relate response variation to the design variables. PLS will also provide you with a response surface visualization of the effects, whenever relevant.

Advanced Topics for Constrained Situations

This section focuses on more technical or "tricky" issues related to the computation of constrained designs.

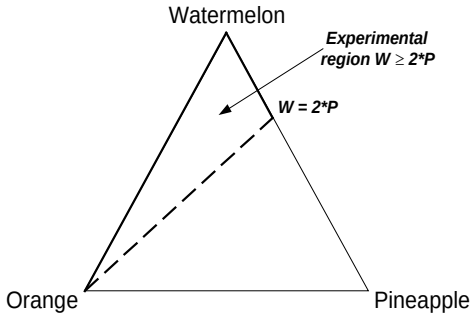
Is the Mixture Region a Simplex?

In a mixture situation where all concentrations vary from 0 to 100%, we have seen in previous chapters that the experimental region has the shape of a simplex. This shape reflects the mixture constraint (sum of all concentrations = 100%).

Note that if some of the ingredients do not vary in concentration, the sum of the mixture components of interest (called Mix Sum in the program) is smaller than 100%, to leave room for the fixed ingredients. For instance if you wish to prepare a fruit punch by blending varying amounts of Watermelon, Pineapple and Orange, with a fixed 10% of sugar, Mix Sum is then equal to 90% and the mixture constraint becomes "sum of the concentrations of all varying components = 90%". In such a case, unless you impose further restrictions on your variables, each mixture component varies between 0 and 90% and ***the mixture region is also a simplex***.

Whenever the mixture components are ***further constrained***, like in the example shown below, the mixture region is ***usually not a simplex***.

With a multi-linear constraint, the mixture region is not a simplex



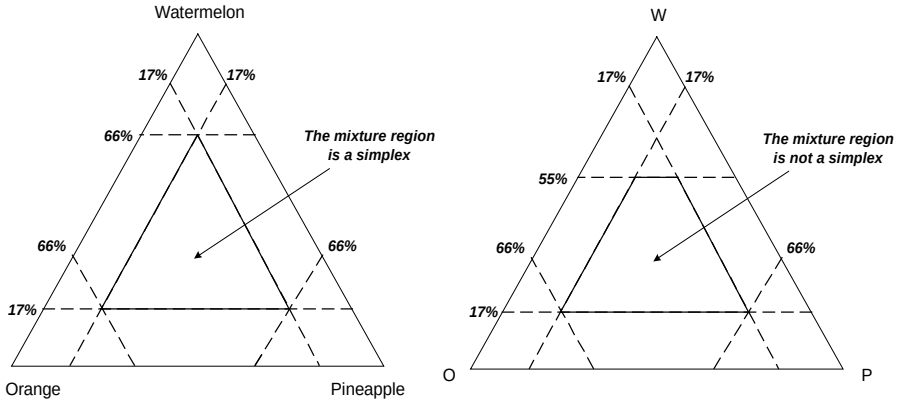
In the absence of Multi-Linear Constraints, the shape of the mixture region depends on the relationship between the lower and upper bounds of the mixture components.

It is a simplex if:

The upper bound of each mixture component is larger than Mix Sum - (sum of the lower bounds of the other components).

The figure below illustrates one case where the mixture region is a simplex and one case where it is not.

Changing the upper bound of Watermelon affects the shape of the mixture region



In the leftmost case, the upper bound of Watermelon is 66% = 100 - (17 + 17): the mixture region is a simplex. If the upper bound of Watermelon is shifted to 0.55, it becomes smaller than 100% - (17 + 17) and the mixture region is no longer a simplex.

Note: When the mixture components only have Lower bounds, the mixture region is always a simplex.

How To Deal with Small Proportions

In a ***mixture situation***, it is important to notice that variations in the major constituents are only marginally influenced by changes in the minor constituents. For instance, an ingredient varying between 0.02 and 0.05% will not noticeably disturb the mixture total; thus it can be considered to vary independently from the other constituents of the blend.

This means that ingredients that are represented in the mixture with a very small proportion can in a way "escape" from the mixture constraint.

So whenever one of the minor constituents of your mixture plays an important role in the product properties, you can investigate its effects by treating it as a process variable. See Chapter How To Combine Mixture and Process Variables p. 38 for more details.

Do You Really Need a Mixture Design?

A special case occurs when all the ingredients of interest have ***small proportions***. Let us consider the following example:

A water-based soft drink consists of about 98% of water, an artificial sweetener, coloring agent, and plant extracts. Even if the sum of the "non-water" ingredients varies from 0 to 3%, the impact on the proportion of water will be negligible.

It does not make any sense to treat such a situation as a true mixture; it will be better addressed by building a classical orthogonal design (full or fractional factorial, central composite, Box-Behnken, depending on your objectives) which focuses on the non-water ingredients only.

How To Select Reasonable Constraints

There are various types of constraints on the levels of design variables. At least three different situations can be considered.

1. Some of the levels or their combinations are ***physically impossible***. For instance: a mixture with a total of 110%, or a negative concentration.
2. Although the combinations are feasible, you know that they are ***not relevant***, or that they will result in ***difficult situations***. Examples: some of the product properties cannot be measured, or there may be discontinuities in the product properties.
3. Some of the combinations that are physically possible and would not lead to any complications are ***not desired***, for instance because of the cost of the ingredients.

When you start defining a new design, think twice about any constraint that you intend to introduce. An unnecessary constraint will not help you solve your problem faster; on the contrary, it will make the design more complex, and may lead to more experiments or poorer results.

Physical constraints

The first two cases mentioned above can be called "real constraints ". You cannot disregard them; if you do, you will end up with missing values in some of your experiments, or uninterpretable results.

Constraints of cost

The third case, however, can be referred to as "imaginary constraints". Whenever you are tempted to introduce such a constraint, examine the impact it will have on the shape of your design. If it turns a perfectly regular and symmetrical situation, which can be solved with a classical design (factorial or classical mixture), into a complex problem requiring a D-optimal algorithm, you will be better off just ***dropping the constraint***.

Build a ***standard*** design, and take the constraint into account afterwards, at the ***result interpretation*** stage. For instance, you can add the constraint to your response surface plot, and select the optimum solution within the constrained region.

This also applies to Upper bounds in mixture components. As mentioned in Chapter “Is the Mixture Region a Simplex?” p.49, if all mixture components have only Lower bounds, the mixture region will automatically be a simplex. Remember that, and avoid imposing an Upper bound on a constituent playing a similar role to the others, just because it is more expensive and you would like to limit its usage to a minimum. It will be soon enough to do this at the interpretation stage, and select the mixture that gives you the desired properties with the smallest amount of that constituent.

How Many Experiments Are Necessary?

In a D-optimal design, the minimum number of experiments can be derived from the shape of the model, according to the basic rule that

In order to fit a model studying p effects, you need at least n=p+1 experiments.

Note that if you stick to that rule without allowing for any extra margin, you will end up with a so-called saturated design, that is to say without any residual degrees of freedom. This is not a desirable situation, especially in an optimization context.

Therefore, The Unscrambler uses the following default number of experiments (**n**), where **p** is the number of effects included in the model:

- For screening designs: n = p + 4 + 3 center samples;
- For optimization designs: n = p + 6 + 3 center samples.

A D-optimal design computed with the default number of experiments will have, in addition to the replicated center samples, enough additional degrees of freedom to provide a reliable and stable estimation of the effects in the model.

However, depending on the geometry of the constrained experimental region, the default number of experiments may not be the ideal one. Therefore, whenever you choose a starting number of points, The

Unscrambler automatically computes 4 designs, with n-1, n, n+1 and n+2 points. The best two are selected and their condition number is displayed, allowing you to choose one of them, or decide to give it another try.

Read more about the choice of a model in Chapter “Relevant Regression Models” in the section about analyzing results from designed experiments, further down in this document.

Three-Way Data: Specific Considerations

If your data consist of two-dimensional spectra (or matrices) for each of your samples, read this chapter to learn a few basics about how these data can be handled in The Unscrambler.

What Is A Three-Way Data Table?

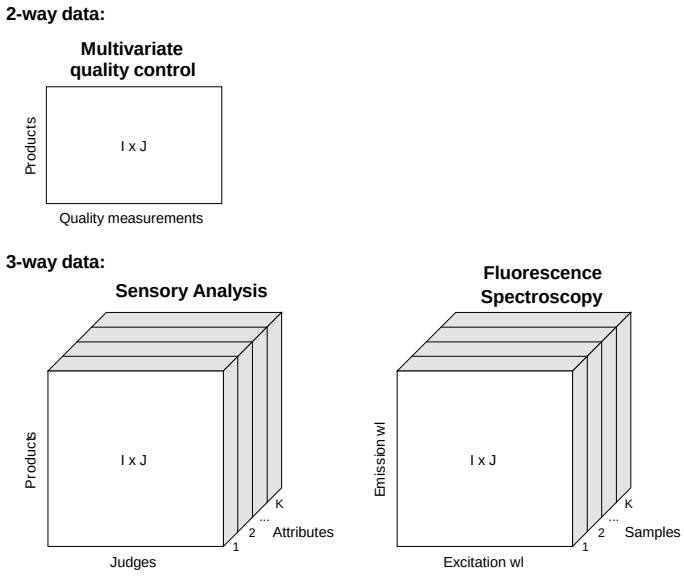
In more and more fields of research and development, the need arises for a relevant way to handle data which do not naturally fit into the classical two-way table scheme.

The figure below illustrates two such cases:

- In ***sensory analysis***, different **products** are rated by several **judges** (or experts, or panelists) using several **attributes** (or ratings, or properties).

- In ***fluorescence spectroscopy***, several **samples** are submitted to an **excitation** light beam at several wavelengths, and respond by **emitting** light, also at several wavelengths.

Examples of two-way and three-way data



Unscrambler users can now import and re-format their three-way data with the help of several new features described in the following sections of this chapter. Before moving on to detailed program operation, let us first define a few useful concepts.

Logical organization Of Three-Way Data Arrays

A classical two-way data table can be regarded as a combination of rows and columns, where rows correspond to **Objects** (samples) and columns to **Variables**.

Similarly, a three-way data array (in The Unscrambler we will simply refer to "3-D data tables") consists of three modes. Most often, one or two of these modes correspond to Objects and the rest to Variables, which leads to two major types of logical organization: "OV^2" and "O^2V".

3D data of type OV^2

One mode corresponds to Objects, while the other two correspond to Variables.

Example: ***Fluorescence spectroscopy***. The Objects are samples analyzed with fluorescence spectroscopy. The Variables are the emission and excitation wavelengths. The values stored in the cells of the 3-D data table indicate the intensity of fluorescence for a given (sample, emission, excitation) triplet.

3D data of type O^2V

Two modes correspond to Objects, while the third one corresponds to Variables.

Example: ***Multivariate image analysis***. The Objects are images consisting of e.g. 256x256 pixels, while the Variables are channels.

OV^2 or O^2V?

Sometimes the difference between the two is subtle and can depend on the question you are trying to answer with your data analysis. Take as an example three-way sensory data, where different products are rated by several judges according to various attributes.

If you consider that usually several samples of the same product are prepared for evaluation by the different judges, and that the results of the assessment of one sample are expressed as a "sensory profile" across the various attributes, then you will clearly choose an O^2V structure for your data. Each sample is a two-way Object determined by a (product, judge) combination, and the Variables are the attributes used for sensory profiling.

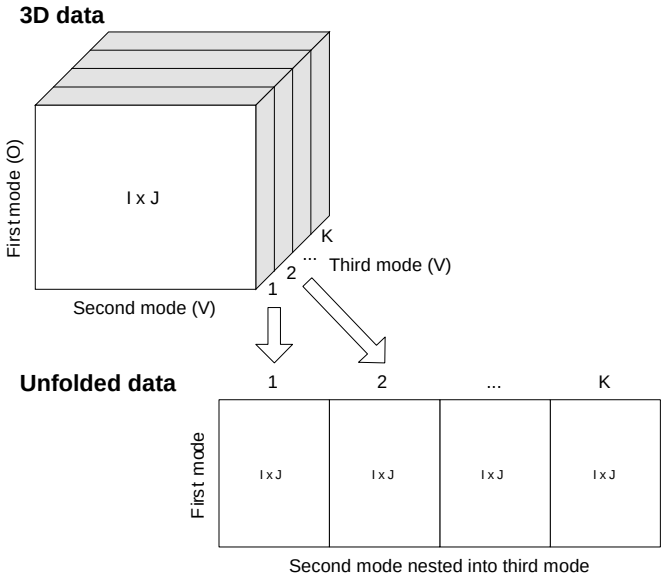
However, if you want to emphasize the fact that each product, as a well-defined Object, can be characterized by the combination of a set of sensory attributes and of individual points of view expressed by the different judges, the data structure reflecting this approach is OV^2.

Unfolding Three-Way Data

Unfolding consists in rearranging a three-way array into a matrix: you take "slices" (or "slabs") of your 3-D data table and put them either on top of each other, or side by side, so as to obtain a "flat" 2-D data table.

The most relevant way to unfold 3-D data is determined by the underlying OV^2 or O^2V structure. The figure below shows the case where the two Variable modes end up as columns of the unfolded table, which has the original Objects as rows. This is the widely accepted way to unfold fluorescence spectra for instance.

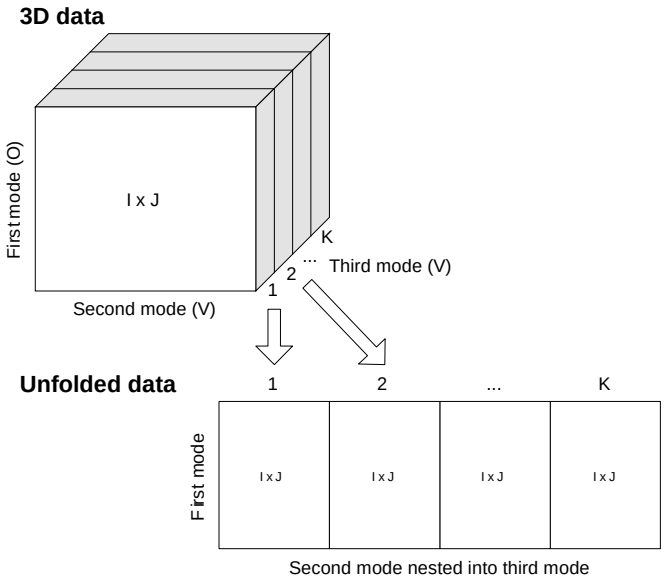
Example: Unfolding an OV^2 array



Primary and Secondary Variables

After unfolding OV^2 data as shown in the figure below, the slabs corresponding to the third mode of the array now form blocks of contiguous columns in the unfolded table. The variables within each block are repeated from block to block with the same layout: the second mode variables have been "nested" into the third mode variables.

Unfolding an OV^2 array

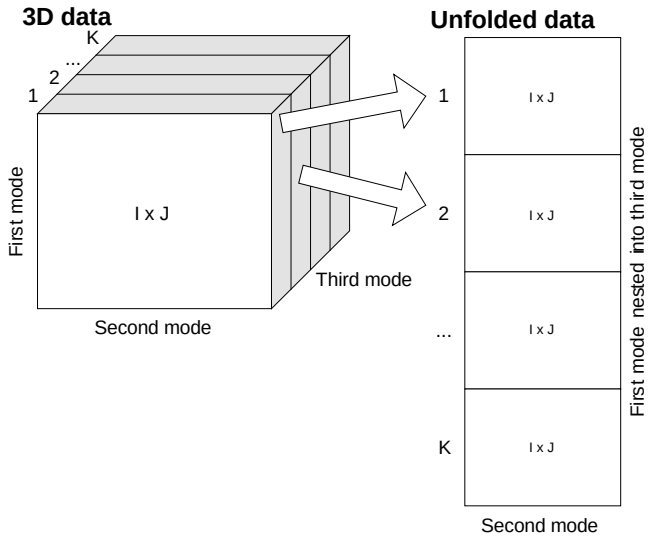


We will call the variables defining the blocks "**primary variables**" (here: k = 1 to K), and the nested variables "**secondary variables**" (here: j = 1 to J).

Primary and Secondary Objects

Let us now imagine that we unfold O^2V data where modes 1 and 3 correspond to the Objects and the second mode to the Variables, and that we rearrange the slabs corresponding to the third mode of the array so that they now form blocks of contiguous rows in the unfolded table (see figure below). The samples within each block are repeated from block to block with the same layout: the first mode samples have been "nested" into the third mode samples.

Unfolding an O^2V array



We will call the samples defining the blocks "**primary samples**" (here: k = 1 to K), and the nested samples "**secondary samples**" (here: i = 1 to I).

Experimental Design and Data Entry in Practice

Menu options and dialogs for experimental design, direct data entry or import from various formats are listed hereafter.

For a detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo's web site www.camo.com/TheUnscrambler/Appendices .

Various Ways To Create A Data Table

The Unscrambler allows you to create new data tables (displayed in an Editor) by way of the following menu options:

- **File - New**;
- **File - New Design**;
- **File - Import**;
- **File - Import 3-D**;
- **File - Convert Vector to Data Table**;
- **File - Duplicate.**

In addition, Drag'n Drop may be used from an existing Unscrambler data table or an external source.

A short description of each menu option follows hereafter. If you need more detailed instructions, read one of the next sections (for instance "Build A Non-designed Data Table" or "Build An Experimental Design") for a list of the commands answering your specific needs.

File - New

The **File - New** option lets you define the size of a new Editor, i.e. the number of samples and variables. It helps you create either a plain 2-D data table, or a 3-D data table with the orientation of your choice. You can then enter the appropriate values in the Editor manually. To name the samples and variables, double-click on the cell where the name is to be displayed and type in the name.

File - New Design

This option takes you into the **Design Wizard**, where you either create a new design or modify or extend an existing one.

File - Import

With the **File - Import** option, you can import a data table from another program. Once you have made all the necessary specifications in the **Import** and **Import from Data Set** dialogs, a new Editor, which contains the imported data, will be created in The Unscrambler.

File - Import 3-D

With the **File - Import 3-D** option, you can import a three-way data table from another program. Once you have made all the necessary specifications in the dialogs, a new Editor, which contains the imported three-way data, will be created in The Unscrambler.

File - Convert Vector to Data Table

This option allows you to create a new data table from a vector, which is especially relevant if the vector is taken from some three-way data.

File - Duplicate

The **File - Duplicate** option contains several choices that allow you to duplicate a designed data table or a three-way data table into a new format. It also allows you to go from a 2-D to a 3-D data structure and vice-versa.

Build A Non-designed Data Table

The menu options listed hereafter allow you to create a new 2-D or 3-D data table, either from scratch or from existing Unscrambler data of various types.

- **File - New…**: Create new 2-D or 3-D from scratch
- **File - Convert Vector to Data Table**: Create new 2-D from a Vector
- **File - Duplicate - As 2-D Data Table**: Create new 2-D from a 3-D
- **File - Duplicate - As 3-D Data Table**: Create new 3-D from a 2-D
- **File - Duplicate - As Non-design**: Create new 2-D from a Design

Build An Experimental Design

The menu options listed hereafter allow you to create a new designed data table, either from scratch or by modifying or extending an existing design.

- **File - New Design**: Create new Design from scratch
- **File - Duplicate - As Modified Design**: Create new Design from existing

Import Data

The menu options listed hereafter allow you to create a new 2-D or 3-D data table by importing from various sources.

- **File - Import**: Import to 2-D
- **File - Import 3-D**: Import to 3-D
- **File - UDI**: Register new DLL for User Defined Import (Supervisor only)

Save Your Data

The menu options listed hereafter allow you to save your data, once you have created a new table or modified it.

- **File - Save**: Save with existing name
- **File - Save As…**: Save with new name

Work With An Existing Data Table

The menu options listed hereafter allow you to open an existing data file, document its properties and close it.

- **File - Open**: Open existing file from browser
- **File - Recent Files List**: Open existing file recently accessed
- **File - Properties**: Document your data and keep log of transformations and analyses
- **File - Close**: Close file

Keep Track Of Your Work With File Properties

Once you have created a new data table, it is recommended to document it: who created it, why, what does it contain? Use **File - Properties** to type in comments in the Notes sheet, and a lot more!

Ready To Work?

Read the next chapters to learn how to make good use of the data in your table:

- Re-formatting and Pre-processing
- Represent Data with Graphs

Then you may proceed by reading about the various methods for data analysis.

Print Your Data

The menu options listed hereafter allow you to print out your data and set printout options.

- **File - Print**: Print out data from the Editor

- **File - Print Preview**: Preview before printout
- **File - Print Lab Report**: Print out randomized list of experiments for your Design
- **File - Print Setup**: Set printout options

Represent Data with Graphs

Principles of graphical data representation and overview of the types of plots available in The Unscrambler.

This chapter presents the graphical tools that facilitate the interpretation of your data and results. You will find a description of all types of plots available in The Unscrambler, as well as some useful tips about how to interpret them.

The Smart Way To Display Numbers

Mean and standard deviation, PCA scores, regression coefficients: All these results from various types of analyses are originally expressed as numbers. Their numerical values are useful, e.g. to compute predicted response values. However, numbers are seldom easy to interpret as such.

Furthermore, the purpose of most of the methods implemented in The Unscrambler is to ***convert numerical data into information***. It would be a pity if numbers were the only way to express this information!

Thus we need an adequate representation of the main results provided by each of the methods available in The Unscrambler. The best way, the most concrete, the one which will give you a real feeling for your results, is the following:

A plot!

Most often, a well-chosen picture conveys a message faster and more efficiently than a long sentence, or a series of numbers. This also applies to your raw data – displaying them in a smart graphical way is already a big step towards understanding the information contained in your numerical data.

However, there are many different ways to plot the same numbers! The trick is to use the most relevant one in each situation, so that the information which matters most is emphasized by the graphical representation of the results.

Different results require different visualizations. This is why there are more than 80 types of predefined plots in The Unscrambler.

The predefined plots available in The Unscrambler can be grouped as belonging to a few different plot types, which are introduced in the next section.

Various Types of Plots

Numbers arranged in a series or a table can have various types of relationships with each other, or be related to external elements which are not explicitly represented by the numbers themselves.

The chosen plot has to reflect this internal organization, so as to give an insight into the structure and meaning of the numerical results.

According to the possible cases of internal relationships between the series of numbers, we can select a graphical representation among six main types of plots:

1. Line plot;
2. 2D scatter plot;
3. 3D scatter plot;

4. Matrix plot;
5. Normal probability plot;
6. Histogram.

In addition, to cover a few special cases, we need two more kinds of representations:

7. Table plot (which is not a plot, as we will see later);
8. Various special plots.

(See Chapter “Special Cases” p.69 for a detailed description of the last two plot types).

Line Plot

A line plot displays a single series of numerical values with a label for each element. The plot has two axes:

- The ***horizontal axis*** shows the ***labels***, in the same physical order as they are stored in the source file;
- The ***vertical axis*** shows the ***scale*** for the plotted numerical values.

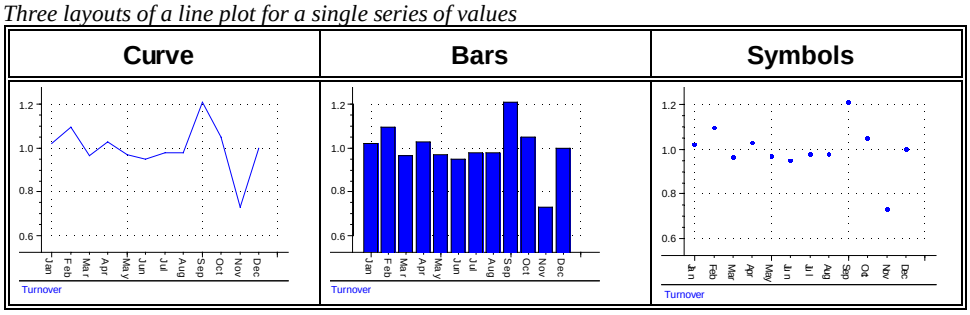
The points in this plot can be represented in several ways:

- A ***curve*** linking the successive points is more relevant if you wish to study a profile, and if the labels displayed on the horizontal axis are ordered in some way (e.g. PC1, PC2, PC3);
- ***Vertical bars*** emphasize the relative size of the numbers;

Symbols produce the same visual impression as a 2D scatter plot (see next chapter

- 2D Scatter Plot), and are therefore not recommended.

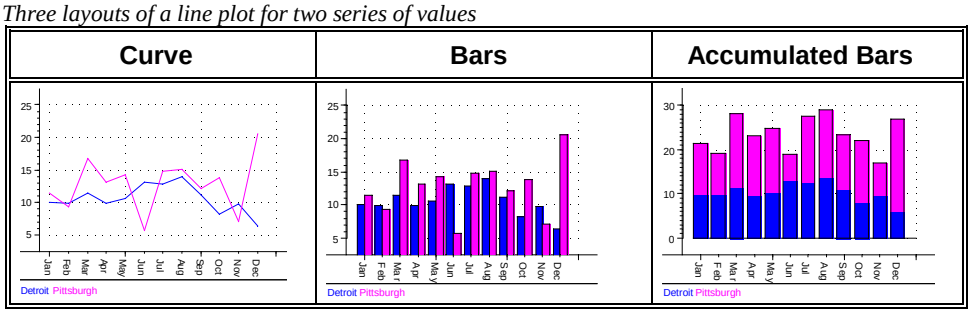
Three layouts of a line plot for a single series of values



Several series of values which share the ***same labels*** can be displayed on the same line plot. The series are then distinguished by means of colors, and an additional layout is possible:

- ***Accumulated bars*** are relevant if the sum of the values for series1, series2, etc... has a concrete meaning (e.g. total production).

Three layouts of a line plot for two series of values



2D Scatter Plot

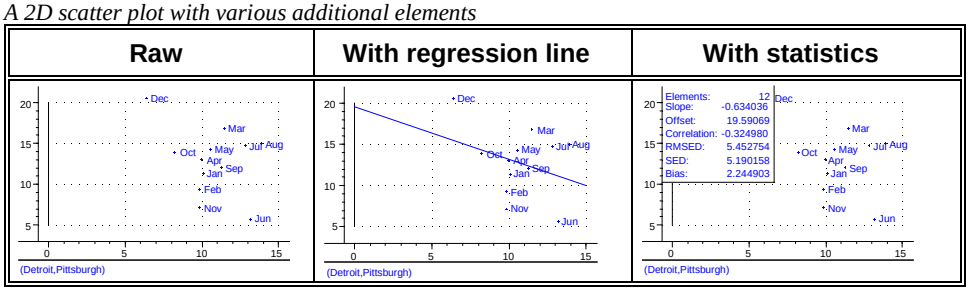
A 2D scatter plot displays two series of values which are related to common elements. The values are shown indirectly, as the coordinates of points in a 2-dimensional space: one point per element.

As opposed to the line plot, where the individual elements are identified by means of a label along one of the axes, ***both axes*** of the 2D scatter plot are used for displaying a numerical ***scale*** (one for each series of values), and the labels may appear beside each point.

Various elements may be added to the plot, to provide more information:

- A ***regression line*** visualizing the relationship between the two series of values;
- A ***target line***, valid whenever the theoretical relationship should be "Y=X";
- ***Plot statistics***, including among others the slope and offset of the regression line (even if the line itself is not displayed) and the correlation coefficient.

A 2D scatter plot with various additional elements



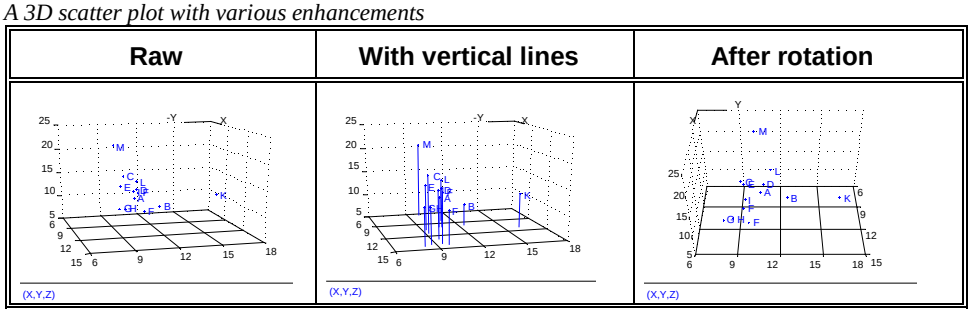
3D Scatter Plot

A 3D scatter plot displays three series of values which are related to common elements. The values are shown indirectly, as the coordinates of points in a 3-dimensional space: one point per element.

3D scatter plots can be enhanced by the following elements:

- ***Vertical lines*** which "anchor" the points can facilitate the interpretation of the plot.
- The plot can be ***rotated*** so as to show the relative positions of the points from a more relevant angle; this can help detect clusters.

A 3D scatter plot with various enhancements



Matrix Plot

The matrix plot can be seen as the 3-dimensional equivalent of a line plot, to display a whole table of numerical values with a label for each element along the 2 dimensions of the table. The plot has up to three axes:

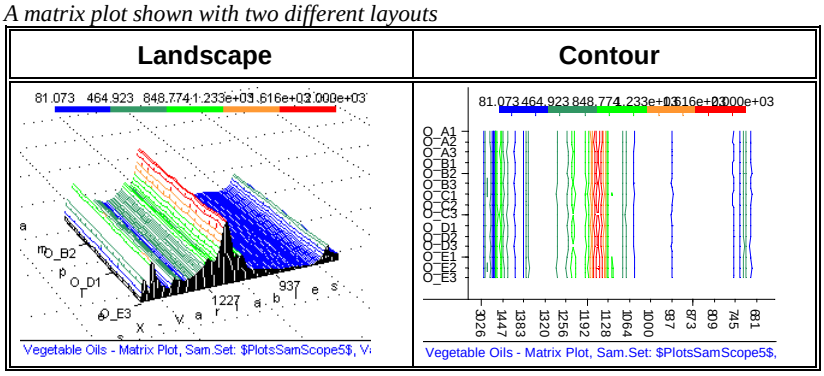
- The ***first two*** show the ***labels***, in the same physical order as they are stored in the source file;
- The ***vertical axis*** shows the ***scale*** for the plotted numerical values.

Depending on the layout, the third axis may be replaced by a color code indicating a range of values.

The points can either be represented individually, or summarized according to one of the following layouts:

- ***Landscape*** shows the table as a 3D surface;
- ***Bars*** give roughly the same visual impression as the landscape plot if there are many points, else the "surface" appears more rugged;
- The ***contour plot*** has only two axes. A few discrete levels are selected, and points (actual or interpolated) with exactly those values are shown as a contour line. It looks like a geographical map with altitude lines;
- On a ***map***, each point of the table is represented by a small colored square, the color depending on the range of the individual value. The result is a completely colored rectangle, where zones sharing close values are easy to detect. The plot looks a bit like an infra-red picture.

A matrix plot shown with two different layouts



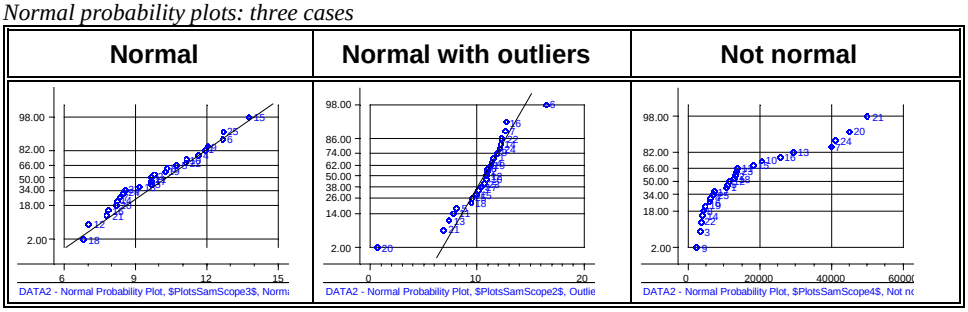
Normal Probability Plot

A normal probability plot displays the cumulative distribution of a series of numbers with a special scale, so that normally distributed values should appear along a straight line. Each element of the series is represented by a point. A label can be displayed beside each point to identify the elements.

This type of plot enables a visual check of the probability distribution of the values:

- If the points are ***close to a straight line***, the distribution is approximately ***normal*** (gaussian);
- If most points are close to a straight line but a few ***extreme values*** (low or high) are far away from the line, these points are ***outliers***;
- If the points are ***not*** close to a ***straight line***, but determine another type of curve, or clusters, the distribution is ***not normal***.

Normal probability plots: three cases


Normal
Normal with outliers
Not normal
DATA2 - Normal Probability Plot, $PlotsSamScope3$, Norma
DATA2 - Normal Probability Plot, $PlotsSamScope2$, Outlie
DATA2 - Normal Probability Plot, $PlotsSamScope4$, Not no

Histogram Plot

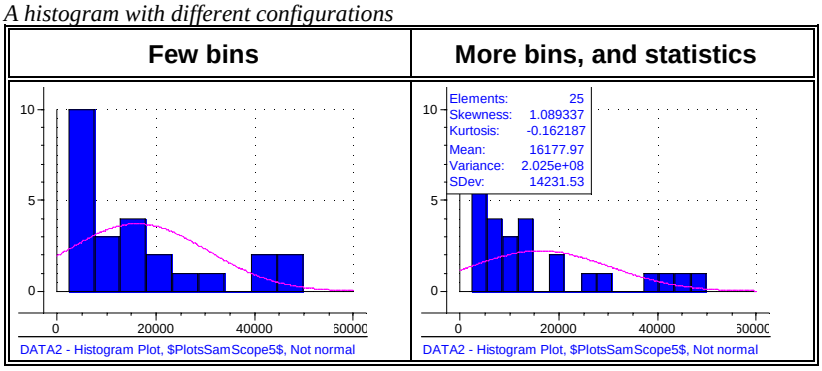
A histogram summarizes a series of numbers without actually showing any of the original elements. The values are divided into ranges (or "***bins***"), and the elements within each bin are counted.

The plot displays the ***ranges*** of values along the ***horizontal*** axis, and the ***number of elements*** as a ***vertical bar*** for each bin.

The graph can be completed by ***plot statistics*** which provide information about the distribution, including mean, standard deviation, skewness (i.e. asymmetry) and kurtosis (i.e. flatness).

It is possible to re-define the ***number of bins***, so as to improve or reduce the smoothness of the histogram.

A histogram with different configurations



Plotting Raw Data

In this section, learn how to plot your data manually from the Editor, using one of the 6 standard types of plots available in The Unscrambler.

Line Plot of Raw Data

Plotting raw data is useful when you want to get acquainted with your data. It is also a necessary element of a data check stage, when you have detected that something is wrong with your data and want to investigate where exactly the problem lies.

Choose a line plot if you are interested in ***individual values***. This is the easiest way to detect which sample has an extreme value, for instance.

- How to do it:

Plot - Line

- How to change plot layout and formatting:

Edit - Options

- How to change plot ranges:

View - Scaling

View - Zoom In

View - Zoom Out

Line Plot of Raw Data: One Row at a Time

This displays values of your variables for a given sample.

Make sure that you select the variables you are interested in. You should also restrict the variable selection to measurements which share a common scale, otherwise the plot might be difficult to read.

Line Plot of Raw Data: Several Rows at a Time

This displays values of your variables for several samples together.

Make sure that you select the variables you are interested in. You should also restrict the variable selection to measurements which share a common scale, otherwise the plot might be difficult to read.

If you have many samples, choose a layout as Curve; it is the easiest to interpret.

Plotting one or several rows of a table as lines is especially useful in the case of spectra: you can see the global shape of the spectrum, and detect small differences between samples.

Line Plot of Raw Data: One Column at a Time

This displays the values of a variable for several samples.

Make sure that you select samples which belong together. If you are interested in studying the structure of the variations from one sample to another, you can sort your table in a special way before plotting the variable. For instance, sort by increasing values of that variable: the plot will show which samples have low values, intermediate values and high values.

Line Plot of Raw Data: Several Columns at a Time

This displays the values of several variables for a set of samples.

Make sure that you select samples which belong together. Also be careful to plot together only variables which share a common scale, otherwise the plot might be difficult to read.

Plotting one or several columns of a table can be a powerful way to display time effects, if your samples have been collected over time. You should then include time information in the table, either as a variable, or implicitly in the sample names, and sort the samples by time before generating the plot.

2D Scatter Plot of Raw Data

Plotting raw data is useful when you want to get acquainted with your data. It is also a necessary element of a data check stage, when you have detected that something is wrong with your data and want to investigate where exactly the problem lies.

Choose a 2D scatter plot if you are interested in the ***relationship between two series of numbers***, their correlation for instance. This is also the easiest way to detect samples which do not comply to the global relationship between two variables.

Since you are usually organizing your data table with samples as rows, and variables as columns, the most relevant 2D scatter plots are those which combine two columns.

Remember to use the specific enhancements to 2D scatter plots if they are relevant:

- Turn on ***Plot Statistics*** if you want to know about the correlation between your two variables;
- Add a ***Regression Line*** if you want to visualize the best linear approximation of the relationship between your two variables;
- Add a ***Target Line*** if this relationship, in theory, is supposed to be "Y=X".
- How to do it:

Plot - 2D Scatter

- How to change plot layout and formatting:

Edit - Options

- How to change plot ranges:

View - Scaling

View - Zoom In

View - Zoom Out

- How to add various elements to a 2D scatter plot:

View - Plot statistics

View - Regression line

View - Target line

3D Scatter Plot of Raw Data

A 3D scatter plot of raw data is most useful when plotting 3 variables, to show the 3-dimensional shape of the swarm of points.

Take advantage of the ***Viewpoint*** option, which rotates the axes of the plot, to make sure that you are looking at your points from the best angle.

- How to do it:

Plot - 3D Scatter

- How to change plot layout and formatting:

Edit - Options

- How to change plot ranges:

View - Scaling

View - Zoom In

View - Zoom Out

- How to change Viewpoint:

View - Rotate

View - Viewpoint - Change

Matrix Plot of Raw Data

A matrix plot of raw data enables you to get an overview of a whole section of your data table. It is especially impressive in its Landscape layout, for spectral data: peaks common to the plotted samples appear as mountains, lower areas of the spectrum build up deep valleys.

Whenever you have a large data table, the matrix plot is an efficient summary. It is mostly relevant, of course, when plotting variables that belong together.

Note: To get a readable matrix plot, select variables measured on the ***same scale***, or sharing a common range of variation.

- How to do it:

Plot - Matrix

Plot - Matrix 3-D

- How to change plot layout and formatting:

Edit - Options

- How to change plot ranges:

View - Scaling

View - Zoom In

View - Zoom Out

- How to change Viewpoint:

View - Rotate

View - Viewpoint - Change

Matrix Plot of Raw Data: Plotting Elements of a Three-Way Data Array

The most relevant way to plot three-way data as a matrix is by selecting a sample (for OV^2 data) or variable (for O^2V) and plot the primary and secondary variables (resp. samples) as a matrix.

Normal Probability Plot of Raw Data

A normal probability plot is the ideal tool for checking whether measured values of a given variable follow a ***normal distribution***. Thus, this plot is most relevant for the columns of your data table. Note that only one column at a time can be plotted.

By extension, if you have reason to believe that your values should be normally distributed, the N-plot also helps you detect ***extreme*** or ***abnormal*** values: they will ***stick out*** either to the top-right or bottom-left of the plot.

- How to do it:

Plot - Normal Probability

- How to change plot layout and formatting:

Edit - Options

- How to add a straight line to a 2D scatter plot:

Edit - Insert Draw Item - Line

Histogram of Raw Data

A histogram is an efficient way to summarize a ***data distribution***, especially for a rather ***large number of values***. In practice, histograms are not relevant for less than 10 values, and start giving you valuable information if you have at least one or two dozen values.

Depending on the context, it can be relevant to plot rows (samples) or columns (variables) as histograms. Like N-plots, histograms can only be obtained for one series of values at a time (one single row or column).

A few special cases are presented in the sections that follow.

- How to do it:

Plot - Histogram

- How to change plot formatting:

Edit - Options

- How to change plot the number of bins:

Edit - Select Bars

- How to add information to your histogram:

View - Plot statistics

- How to transform your data:

Modify - Compute General

Histogram of Raw Data: Detecting the Need for a Transformation

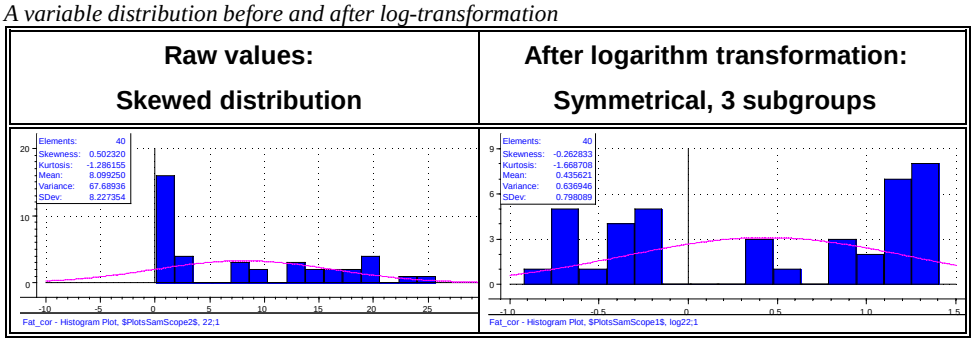
Multivariate analyses, linear regression and ANOVA have one assumption in common: relationships between variables can be summarized using ***straight lines*** (to put it simply). This implies that the models will only perform reliably if the data are ***balanced***.

This assumption is violated for data with ***skewed*** (asymmetrical) distributions: there is more weight at one end of the range of variation than at the opposite end.

If your analysis contains variables with heavily skewed distributions, you run the risk that some samples, lying at the "tail" of the distribution, will be considered outliers. This is a wrong diagnosis: Something is the matter with the whole distribution, not with a single value.

In such cases, it is recommended to implement a ***transformation*** that will make the distribution more balanced. Whenever you have a positive skewness, which is the most often encountered case, a logarithm usually fixes the problem, as shown hereafter.

A variable distribution before and after log-transformation



Note: There is nothing wrong with a non-normal distribution in itself. There can be 3 balanced groups of values, "low", "medium" and "high". Only highly ***skewed distributions*** are dangerous for multivariate analyses.

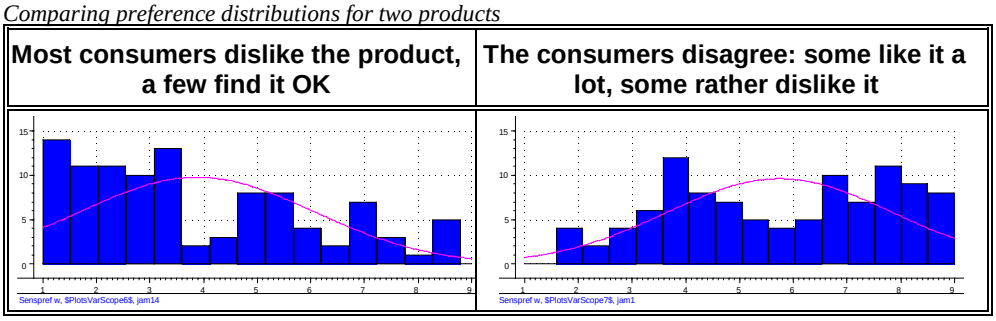
Histogram of Raw Data: Preference Ratings

Preference ratings from a consumer study where other types of data have also been collected, can be delicate to handle in a classical way. If you are studying several products, and want to check how well your many consumers agree on their ratings, you cannot directly summarize your data with the classical plots available for descriptive statistics (percentiles, mean and standard deviation) because your products are stored as rows of your data table, and each consumer builds up a column (variable).

Unless you want to start some manipulations involving the selection of a fraction of your data table and a transposition, the simple and efficient way to summarize the preference ratings for a given product (before starting a multivariate analysis) is to plot row histograms.

Look for groups of consumers with similar ratings: very often, ***subgroups*** are more interesting than the "average" opinion!

Comparing preference distributions for two products


Most consumers dislike the product, a few find it OK
The consumers disagree: some like it a lot, some rather dislike it
Senspref w, $PlotsVarScope6$, jam14
Senspref w, $PlotsVarScope7$, jam1

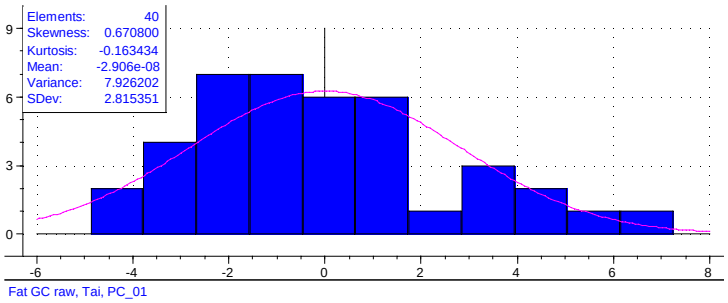
Note: Configure your histograms with a relevant number of bars, to get enough details.

Histogram of Raw Data: Plot Results as a Histogram

Although there is no ***predefined*** histogram plot of analysis results, it is possible to plot any kind of results as a histogram by taking advantage of the **Results - General View** command.

This is how, for instance, you can check whether your samples are ***symmetrically distributed*** on a score plot. shows an example where the scores along PC1 have a skewed distribution: It is likely that several of the variables taken into account in the analysis require a logarithm transformation.

Histogram of PCA scores



Special Cases

This section presents a few types of graphical data representations which do not fit in any of the 6 standard plot types described in Chapter Various Types of Plots. These types of plots are ***not*** available for manual plotting of raw data from the Editor.

Special Plots

This is an ad-hoc category which groups all plots that do not fit into any of the other descriptions.

Some are an adaptation of existing plot types, with an additional enhancement. For instance, “Means” can be displayed as a line plot; if you wish to include standard deviations (SDev) into the same plot, the most relevant way to do so is to

1. configure the plot layout as ***bars***;
2. and display SDev as an ***error bar*** on top of the Mean vertical bar.

This is what has been done in the special plot “Mean and Sdev”.

Other special plots have been developed to answer specific needs, e.g. visualize the outcome of a Multiple Comparisons test in a graphical way which gives immediate overview.

Two examples of special plots

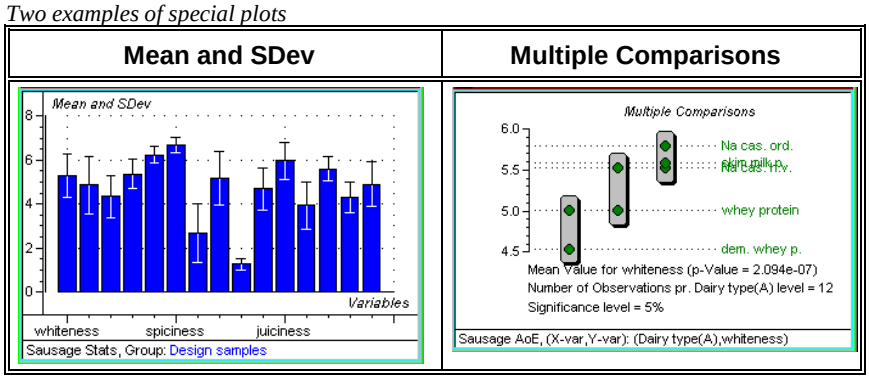


Table Plot

A table plot is nothing but results arranged in a table format, displayed in a graphical interface which optionally allows for re-sizing and sorting of the columns of the table. Although it is not a “plot” as such, it allows tabulated results to be displayed in the same Viewer system as other plots.

The table plot format is used under two different circumstances:

1. A few analysis results require this format, because it is the only way to get an interpretable ***summary of complex results***. A typical example is Analysis of Variance (ANOVA); some of its individual results can be plotted separately as line plots, but the only way to get a full overview is to study 4 or 5 columns of the table simultaneously.
2. Standard graphical plots like line plots, 2D scatter plots, matrix plots… can be displayed numerically to facilitate the exportation of the underlying numbers to another graphical package, or a worksheet.

Two different types of table plots

Effects Overview

Effects Overview
Significance Testing Method: HOIE

Variable	Glossy	Shape	Adh
Add DM(A)	NS	NS	NS
pH(B)	NS	-	NS
Dry Matter(C)	---	++	NS
Maturity(D)	NS	+	--
AB	NS	NS	NS
AC	NS	NS	NS
AD	NS	NS	NS
BC	--	+	-
BD	-	NS	-
CD	NS	NS	NS

Analysis of eff...

Numerical view of a plot

	ResXCalTot (PCs)	ResXValTot (PCs)
PC 00	0.917	1.091
PC 01	0.445	0.915
PC 02	0.210	0.851
PC 03	7.270e-02	0.647
PC 04	3.231e-03	0.533
PC 05	5.973e-04	1.021
PC 06	2.403e-16	m

RESULT4, Variable: *c.Total v.Total*

- How to display a plot as numbers:

View - Numerical

Re-formatting and Pre-processing

This chapter focuses on all the operations that change the layout or the values in your data table.

What Is Re-formatting?

Changing the layout of a data table is called ***re-formatting***.

Here are a few examples:

1. Get a better overview of the contents of your data table by ***sorting*** variables or samples.

2. Change point of view: by ***transposing*** a data table, samples become variables and vice-versa.

3. Apply a 2-D analysis method to 3-D data: by ***unfolding*** a three-way data array, you enable the use of e.g. PCA on your data.

What Is Pre-processing?

Introducing changes in the values of your variables, e.g. so as to make them better suited for an analysis, is called ***pre-processing***. One may also talk about applying a ***pre-treatment*** or a ***transformation***.

Here are a few examples:

1. Improve the distribution of a skewed variable by taking its ***logarithm***.

2. Remove some noise in your spectra by ***smoothing*** the curves.

3. Improve the precision in your sensory assessments by taking the ***average*** of the sensory ratings over all panelists.

4. Allow plotting of all raw data and use of "classical" analysis methods by ***filling missing*** values with values estimated from the non-missing data.

Other operations

In addition, section Make Simple Changes In The Editor shows you how to perform various editing operations like adding new samples or variables, or creating a Category variable.

Principles of Data Pre-processing

In this chapter, read about how to make your data better suited for a specific analysis.

A wide range of ***transformations*** can be applied to data before they are analyzed. The main purpose of transformations is to make the distribution of given variables more suitable for a powerful analysis.

The sections that follow detail the various types of transformations available in The Unscrambler.

Sometimes it may also be necessary to change the layout of a data table so that a given transformation or analysis becomes more relevant. This is the purpose of ***re-formatting***.

Finally, a number of simple ***editing operations*** may be required:

- in order to ***improve the interpretation*** of future results (e.g. insert a category variable whose levels describe the samples in your table qualitatively);
- as a ***safety measure*** (e.g. make a copy of a variable before you take its logarithm);

- as a ***pre-requisite*** before the desired re-formatting or transformation can be applied (e.g. create a new column where you can compute the ratio of two variables).

Re-formatting and editing operations will not be described in detail here; you may lookup the specific operation you are interested in by checking section Re-formatting and Pre-processing in Practice.

Filling Missing Values

It may sometimes be difficult to gather values of all the variables you are interested in, for all the samples included in your study. As a consequence, some of the cells in your data table will remain empty. This may also occur if some values are lost due to human or instrumental failure, or if a recorded value appears so improbable that you have to delete it, thus creating an empty cell.

Using the **Edit - Fill Missing** menu option from the Data Editor, you can fill those cells with values estimated from the information contained in the rest of the data table.

Although some of the analysis methods (PCA, PCR, PLS, MCR) available in The Unscrambler can cope with a reasonable amount of missing values, there are still multiple advantages in filling empty cells with estimated values:

- Allow all points to appear on a 2-D or 3-D scatter plot;
- Enable the use of transformations requiring that all values are non-missing, like for instance derivatives;
- Enable the use of analysis methods requiring that all values are non-missing, like for instance MLR or Analysis of Effects.

Two methods are available for the estimation of missing values:

- ***Principal Component Analysis*** performs a reconstruction of the missing values based on a PCA model of the data with an optimal number of components. This fill missing procedure is the *default* selection and the recommended method of choice for ***spectroscopic*** data.
- ***Row Column Mean Analysis*** only makes use of the *same column and row* as each cell with missing data. Use this method if the columns or rows in your data come from very different sources that do not carry information about other rows or columns. This can be the case for process data.

Computation of Various Functions

Using the **Modify - Compute General** function from the Data Editor, you can apply any kind of function to the vectors of your data matrices (or to a whole matrix).

One of the most widely used is the ***logarithmic transformation***, which is especially useful to make the distribution of skewed variables more symmetrical. It is also indicated when the measurement error on a variable increases proportionally with the level of that variable; taking the logarithm will then achieve uniform precision over the whole range of variation. This particular application is called ***variance stabilization***.

In cases of only slight asymmetry, a ***square root*** can serve the same purposes as a logarithm.

To decide whether some of your data require such a transformation, plot a histogram of your variables to investigate their ***distribution***.

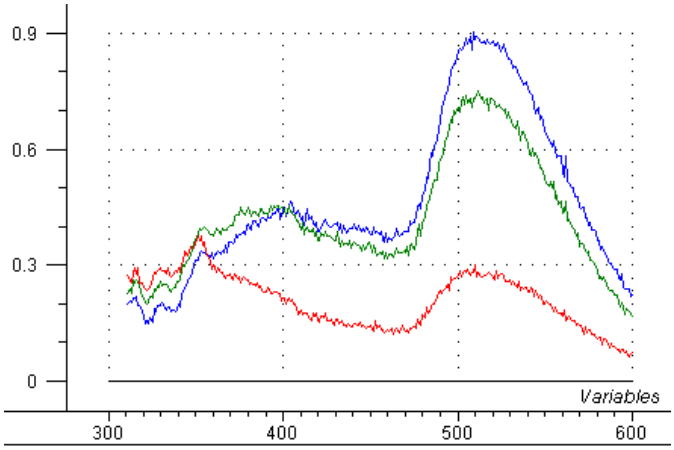
Smoothing

This transformation is relevant for variables which are themselves a function of some underlying variable, for instance time, or in the existence of intrinsic spectral intervals.

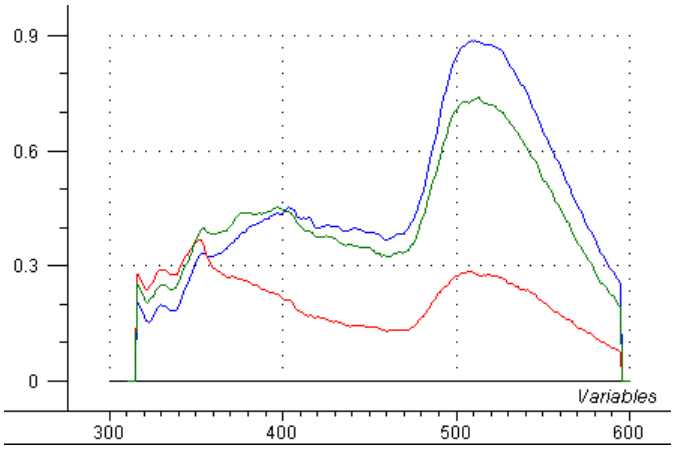
In The Unscrambler, you have the choice between four smoothing algorithms:

1. ***Moving average*** is a classical smoothing method, which replaces each observation with an average of the adjacent observations (including itself). The number of observations on which to average is the user-chosen "segment size" parameter.
2. ***Savitzky-Golay***
   The Savitzky-Golay algorithm fits a polynomial to each successive curve segment, thus replacing the original values with more regular variations. You can choose the length of the smoothing segment (or right and left points separately) and the order of the polynomial. It is a very useful method to effectively remove spectral noise spikes while chemical information can be kept, as shown in the figures below.

Raw UV / Vis spectra show noise spikes



UV / Vis spectra after Savitzky-Golay smoothing at 11 smoothing points and 2nd polynomial degree setting



3. ***Median filtering*** replaces each observation with the median of its neighbors. The number of observations from which to take the median is the user-chosen "segment size" parameter; it should be an odd number.
4. ***Gaussian filtering*** is a weighted moving average where each point in the averaging function is affected a coefficient determined by a Gauss function with $\sigma^2 = 2$. The further away the neighbor is, the smaller the coefficient, so that information carried by the smoothed point itself and its nearest neighbors is given greater importance than in an un-weighted moving average.

Example:

Let us compare the coefficients in a Moving average and a Gaussian filter for a data segment of size 5.

If the data point to be smoothed is x_k, the segment consists of the 5 values x_{k-2}, x_{k-1}, x_k, x_{k+1} and x_{k+2}.

The Moving average is computed as:

$(x_{k-2} + x_{k-1} + x_k + x_{k+1} + x_{k+2})/5$

that is to say

$0.2*x_{k-2} + 0.2*x_{k-1} + 0.2*x_k + 0.2*x_{k+1} + 0.2*x_{k+2}$

The Gaussian distribution function for a 5-point segment is:

0.0545 0.2442 0.4026 0.2442 0.0545

As a consequence, the Gaussian filter is:

$0.0545*x_{k-2} + 0.2442*x_{k-1} + 0.4026*x_k + 0.2442*x_{k+1} + 0.0545*x_{k+2}$

As you can see, points closer to the center have a larger coefficient in the Gaussian filter than in the moving average, while the opposite is true of points close to the borders of the segment.

Normalization

Normalization is a family of transformations that are computed sample-wise. Its purpose is to "scale" samples in order to achieve specific properties.

The following normalization methods are available in The Unscrambler:

1. Area normalization;
2. Unit vector normalization;
3. Mean normalization;
4. Maximum normalization;
5. Range normalization;
6. Peak normalization.

Area Normalization

This transformation normalizes a spectrum X_i by calculating the area under the curve for the spectrum. It attempts to correct the spectra for indeterminate path length when there is no way of measuring it, or isolating a band of a constant constituent.

$$newX_i = X_i / \sum_j x_{i,j}$$

Property of area-normalized samples:

The area under the curve becomes the same for all samples.

Note: In practice, area normalization and mean normalization (see Mean Normalization) only differ by a constant multiplicative factor. The reason why both are available in The Unscrambler is that, while **spectroscopists** may be more familiar with ***area normalization***, other groups of users may consider ***mean normalization*** a more "standard" method.

Unit vector Normalization

This transformation normalizes sample-wise data X_i to unit vectors. It can be used for pattern normalization, which is useful for pre-processing in some pattern recognition applications.

$$newX_i = X_i / SQRT(\sum_j x_{i,j}^2)$$

Property of unit vector normalized samples:

The normalized samples have a length (“norm”) of 1.

Mean Normalization

This is the most classical case of normalization.

It consists in ***dividing each row*** of a data matrix by its ***average***, thus neutralizing the influence of the hidden factor.

It is equivalent to replacing the original variables by a profile centered around 1: only the relative values of the variables are used to describe the sample, and the information carried by their absolute level is dropped. This is indicated in the specific case where all variables are measured in the same unit, and their values are assumed to be proportional to a factor which cannot be directly taken into account in the analysis.

For instance, this transformation is used in ***chromatography*** to express the results in the same units for all samples, no matter which volume was used for each of them.

Caution! This transformation is not relevant if all values of the curve do not have ***the same sign***. It was originally designed for positive values only, but can easily be applied to all-negative values through division by the absolute value of the average instead of the raw average. Thus the original sign is kept.

Property of mean-normalized samples:

The area under the curve becomes the same for all samples.

Maximum Normalization

This is an alternative to classical normalization which ***divides each row*** by its ***maximum absolute value*** instead of the average.

Caution! The relevance of this transformation is doubtful if all values of the curve do not have ***the same sign***.

Property of maximum-normalized samples:

- If all values are positive: the maximum value becomes +1.
- If all values are negative: the minimum value becomes -1.
- If the sign of the values changes over the curve: either the maximum value becomes +1 or the minimum value becomes -1.

Range Normalization

Here each row is ***divided by its range***, i.e. max value - min value.

Property of range-normalized samples:

The curve span becomes 1.

Peak Normalization

This transformation normalizes a spectrum X_i by the chosen k^{th} spectral point, which is always chosen for both training set and "unknowns" for prediction.

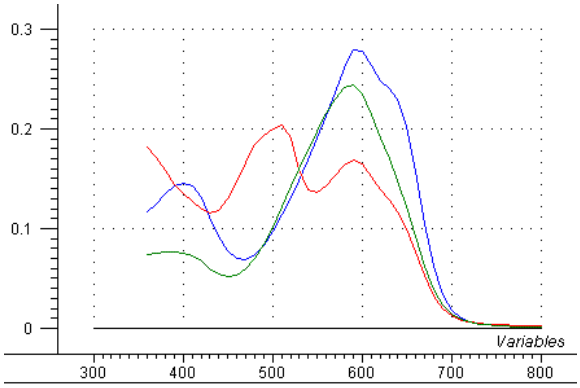
$$newX_i = X_i / x_{i,k}$$

It attempts to correct the spectra for indeterminate path length. Since the chosen spectral point (usually the maximum peak of a band of the constant constituent, or the isosbestic point) is assumed to be concentration invariant in all samples, an increase or decrease of the point intensity can be assumed to be entirely due to an increase or decrease in the sample path length. Therefore, by normalizing the spectrum to the intensity of the peak, the path length variation is effectively removed.

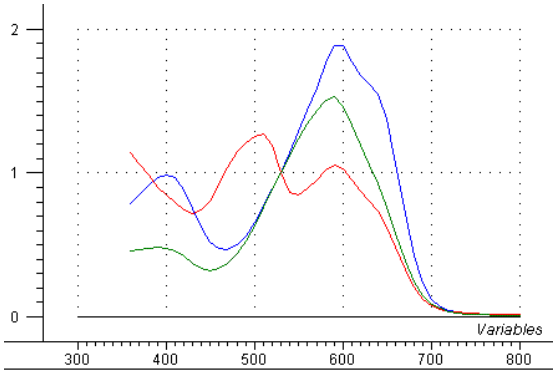
Property of peak-normalized samples:

All transformed spectra take value 1 at the chosen constant point, as shown in the figures below.

Raw UV / Vis spectra



Spectra after peak normalization at 530 nm, the isosbestic point



Caution! One potential problem with this method is that it is extremely susceptible to baseline offset, slope effects and wavelength shift in the spectrum.

The method requires that the samples have an isosbestic point, or have a constant concentration constituent and that an isolated spectral band can be identified which is solely due to that constituent.

Spectroscopic Transformations

Specific transformations for spectroscopy data are simply a ***change of units***.

The following transformations are possible:

- ***Reflectance to absorbance***,
- ***Absorbance to reflectance***,
- ***Reflectance to Kubelka-Munk***.

Multiplicative Scatter Correction

(MSC / EMSC)

Multiplicative Scatter Correction (MSC) is a transformation method used to compensate for additive and/or multiplicative effects in spectral data. Extended Multiplicative Scatter Correction (EMSC) works in a similar way; in addition, it allows for compensation of wavelength-dependent spectral effects.

MSC

MSC was originally designed to deal with multiplicative scattering alone. However, a number of similar effects can be successfully treated with MSC, such as:

- path length problems,

- offset shifts,

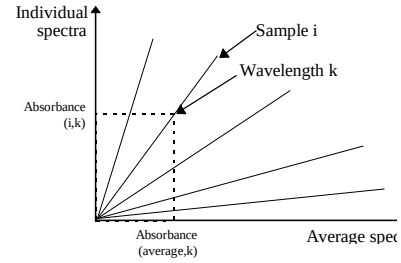
- interference, etc.

The idea behind MSC is that the two effects, ***amplification*** (multiplicative) and ***offset*** (additive), should be ***removed*** from the data table to avoid that they dominate the information (signal) in the data table.

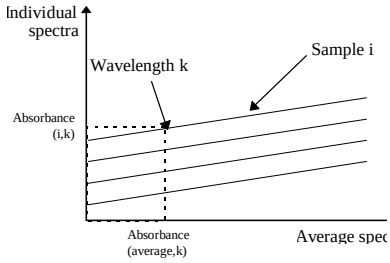
The correction is done by two simple transformations. Two correction coefficients, *a* and *b*, are calculated and used in these computations, as represented graphically below:

Multiplicative Scatter Correction

Multiplicative Scatter Effect:



Additive Scatter Effect:



The correction coefficients are computed from a regression of each individual spectrum onto the average spectrum. Coefficient *a* is the intercept (offset) of the regression line, coefficient *b* is the slope.

EMSC

EMSC is an extension to conventional MSC, which is not limited to only removing multiplicative and additive effects from spectra. This extended version allows a separation of physical light scattering effects from chemical light absorbance effects in spectra.

In EMSC, new parameters *h*, *d* and *e* are introduced to account for physical and chemical phenomena that affect the measured spectra. Parameters *d* and *e* are wavelength specific, and used to compensate regions where such unwanted effects are present. EMSC can make estimates of these parameters, but the best result is obtained by providing prior knowledge in form of spectra that are assumed to be relevant for one or more of the underlying constituents within the spectra and spectra containing undesired effects. The parameter *h* is estimated on the basis of a reference spectrum representative for the data set, either provided by the user or calculated as the average of all spectra.

Adding Noise

Contrary to the other transformations, adding noise to your data would seem to decrease the precision of the analysis.

This is exactly the purpose of that transformation: Include some ***additive*** or ***multiplicative*** noise in the variables, and see how this ***affects the model***.

Use this option only when you have modeled your original data satisfactorily, to check how well your model may perform if you use it for future predictions based on new data assumed to be more noisy than the calibration data.

Derivatives

Like smoothing, this transformation is relevant for variables which are themselves a function of some underlying variable, e.g. absorbance at various wavelengths. Computing a derivative is also called differentiation.

In The Unscrambler, you have the choice among three methods for computing derivatives, as described hereafter.

Savitzky-Golay Derivative

Enables you to compute 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order derivatives. The Savitzky-Golay algorithm is based on performing a least squares linear regression fit of a polynomial around each point in the spectrum to smooth the data. The derivative is then the derivative of the fitted polynomial at each point. The algorithm includes a smoothing factor that determines how many adjacent variables will be used to estimate the polynomial approximation of the curve segment.

Gap-Segment Derivative

Enables you to compute 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order derivatives. The parameters of the algorithm are a gap factor and a smoothing factor that are determined by the segment size and gap size chosen by the user.

The principles of the Gap-Segment derivative can be explained shortly in the simple case of a 1^{st} order derivative. If the function y=f(x) underlying the observed data varies slowly compared to sampling frequency, the derivative can often be approximated by taking the difference in y-values for x-locations separated by more than one point. For such functions, Karl Norris suggested that derivative curves with less noise could be obtained by taking the difference of two averages, formed by points surrounding the selected x-locations. As a further simplification, the division of the difference in y-values, or the y-averages, by the x-separation Δx, is omitted.

Norris introduced the term ***segment*** to indicate the length of the x-interval over which y-values are averaged, to obtain the two values that are subtracted to form the estimated derivative.

The ***gap*** is the length of the x-interval that separates the two segments that are averaged.

You may read more about Norris derivatives (implemented as Gap-Segment and Norris-Gap in The Unscrambler) in Hopkins DW, What is a Norris derivative?, *NIR News* Vol. 12 No. 3 (2001), 3-5. See chapter Method References for more references on derivatives.

Norris-Gap Derivative

It is a special case of Gap-Segment Derivative with segment size = 1.

Property of Gap-segment and Norris-Gap Derivatives:

Dr. Karl Norris has developed a powerful approach in which two distinct items are involved. The first is the Gap Derivative, the second is the "Norris Regression", which may or may not use the derivatives. The applications of the Gap Derivative are to improve the rejection of interfering absorbers. The "Norris Regression" is a regression procedure to remove the impact of varying path lengths among samples due to scatter effects.

More About Derivative Methods and Applications

Derivative attempts to correct for baseline effects in spectra for the purpose of creating robust calibration models.

1st Derivative

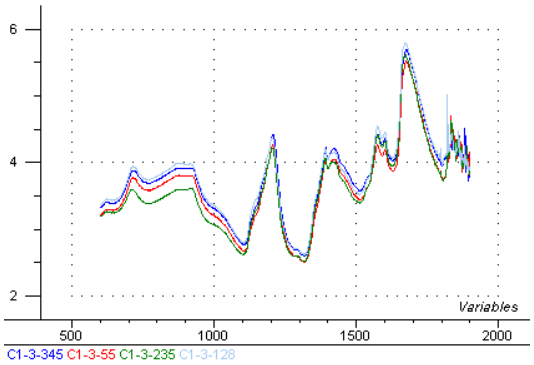
The 1st derivative of a spectrum is simply a measure of the slope of the spectral curve at every point. The slope of the curve is not affected by baseline offsets in the spectrum, and thus the 1st derivative is a very effective method for removing baseline offsets. However, peaks in raw spectra usually become zero-crossing points in 1st derivative spectra, which can be difficult to interpret.

Example:

Public NIR transmittance spectra for an active pharmaceutical ingredient (API) recorded in the range of 600-1980 nm in 2 nm increments. API = 175.5 for spectra C1-3-345 and C1-3-55; API = 221.5 for spectra C1-3-235 and C1-3-128.

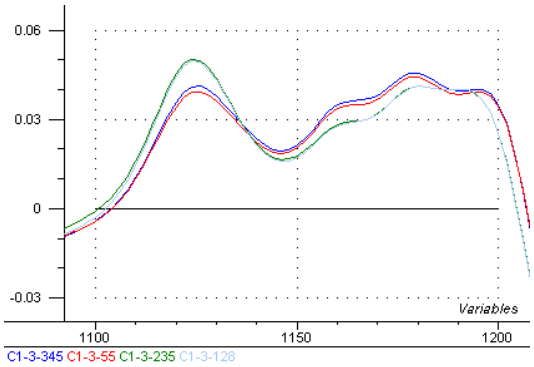
The figure below shows severe baseline offsets and possible linear tilt problems, and two levels of API spectra are not separated.

Public NIR transmittance spectra for an active pharmaceutical ingredient (API) recorded in the range of 600-1980 nm in 2 nm increments: raw spectra



The next figure displays the 1st order derivative spectra at the region of 1100-1200 nm (Savitzky-Golay derivative, 11 points segment and 2nd order of polynomial). One can see the baseline offsets effectively removed, and spectra of two levels of API separated. Note that a peak around 1206 nm crosses zero.

1st order derivative spectra at the region of 1100-1200 nm.



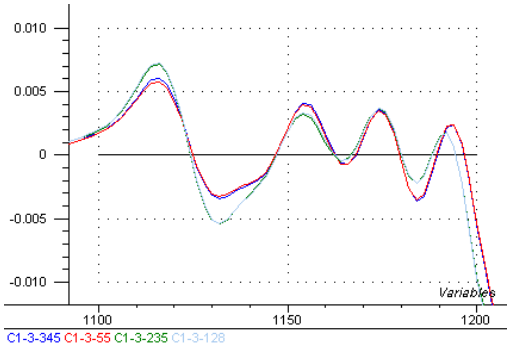
2nd Derivative

The 2nd derivative is a measure of the change in the slope of the curve. In addition to ignoring the offset, it is not affected by any linear "tilt" that may exist in the data, and is therefore a very effective method for removing both the baseline offset and slope from a spectrum. The 2nd derivative can help resolve nearby peaks and sharpen spectral features. Peaks in raw spectra usually change sign and turn to negative peaks.

Example:

On the same data as in the previous example, a 2nd order derivative has been computed at the region of 1100-1200 nm (Savitzky-Golay derivative, 11 points segment and 2nd order of polynomial). One can see the spectra of two levels of API separated, as well as overlapped spectral features enhanced.

2nd order derivative spectra at the region of 1100-1200 nm.



3rd and 4th Derivatives

3rd and 4th derivatives are available in the Unscrambler although they are not as popular as 1st and 2nd derivatives. They may reveal phenomena which do not appear clearly when using lower-order derivatives.

Savitzky-Golay vs. Gap-Segment

The Savitzky-Golay method and the Gap-Segment method use information from a localized segment of the spectrum to calculate the derivative at a particular wavelength rather than the difference between adjacent data points. In most cases, this avoids the problem of noise enhancement from the simple difference method and may actually apply some smoothing to the data.

The Gap-Segment method requires gap size and smoothing segment size (usually measured in wavelength span, but sometimes in terms of data points). The Savitzky-Golay method uses a convolution function, and thus the number of data points (segment) in the function must be specified. If the segment is too small, the result may be no better than using the simple difference method. If it is too large, the derivative will not

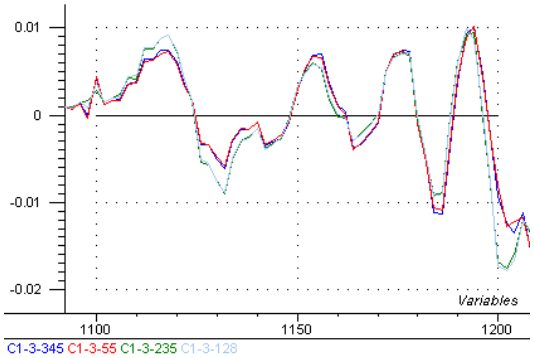
represent the local behaviour of the spectrum (especially in the case of Gap-Segment), and it will smooth out too much of the important information (especially in the case of Savitzky-Golay). Although there have been many studies done on the appropriate size of the spectral segment to use, a good general rule is to use a sufficient number of points to cover the full width at half height of the largest absorbing band in the spectrum. One can also find optimum segment sizes by checking model accuracy and robustness under different segment size settings.

Example:

The data are still the same as in the previous examples.

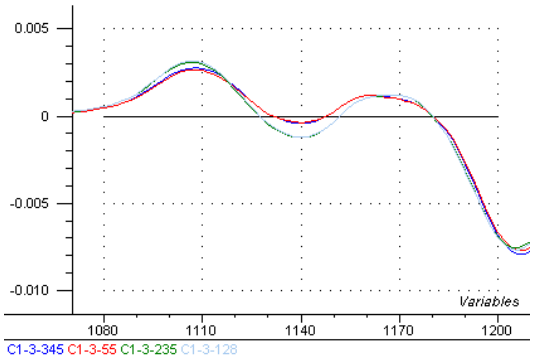
In the next figure, you can see what happens when the selected segment size is too small (Savitzky-Golay derivative, 3 points segment and 2nd order of polynomial). One can see noisy features in the region.

Segment size is too small: 2nd order derivative spectra at the region of 1100-1200 nm.



In the figure that follows, the selected segment size is too large: (Savitzky-Golay derivative, 31 points segment and 2nd order of polynomial). One can see that some relevant information has been smoothed out.

Segment size is too large: 2nd order derivative spectra at the region of 1100-1200 nm.



The main disadvantage of using derivative pre-processing is that the resulting spectra are very difficult to interpret. For example, the PLS loadings for the calibration model represent the changes in the constituents of interest. In some cases (especially in the case of PLS-1 models), the loadings can be visually identified as representing a particular constituent. However, when derivative spectra are used, the loadings cannot be easily identified. A similar situation exists in regression coefficient interpretation. In addition, the derivative makes visual interpretation of the residual spectrum more difficult, so that for instance finding spectral location for impurities in the samples cannot be done.

Standard Normal Variate

Standard Normal Variate (SNV) is a row-oriented transformation which centers and scales individual spectra.

Each value in a row of data is transformed according to the formula:

New value = (Old value – mean (Old row)) / Stdev (Old row)

Like MSC (see Multiplicative Scatter Correction), the practical result of SNV is that it removes scatter effects from spectral data.

An effect of SNV is that on the vertical scale, each spectrum is centered on zero and varies roughly from –2 to +2. Apart from the different scaling, the result is similar to that of MSC. The practical difference is that SNV standardises each spectrum using only the data from that spectrum; it does not use the mean spectrum of any set. The choice between SNV and MSC is a matter of taste.

Averaging

Averaging over samples (in case of ***replicates***) or over variables (for ***variable reduction***, e.g. to reduce the number of spectroscopic variables) may have, depending on the context, the following advantages:

- Increase precision;
- Get more stable results;
- Interpret the results more easily.

Application example:

Improve the precision in your ***sensory assessments*** by taking the average of the sensory ratings over all panelists.

Transposition

Matrix transposition consists in ***exchanging rows for columns*** in the data table.

It is particularly useful if the data have been imported from external files where they were stored with one row for each variable.

Shifting Variables

Shifting variables is much used on ***time-dependent*** data, such as for processes where the output measurement is ***time-delayed*** relative to input measurements.

To make a meaningful model of such data you have to shift the variables so that each row contains "synchronized" measurements for each sample.

User-Defined Transformations

The transformation that your specific type of data requires may not be included as a predefined choice in The Unscrambler. If this is the case, you have the possibility to ***register your own transformation*** for use in the Unscrambler as User-Defined Transformation (UDT).

Such transformation components have to be developed separately (e.g. in Matlab), and installed on the computer when needed. A wide range of modifications can be done by such components, including deleting and inserting both variables and samples.

You may register as many UDTs as you wish.

Centering

As a rule, the first stage in multivariate modeling using projection methods is to subtract the average from each variable. This operation, called ***mean-centering***, ensures that all results will be interpretable in terms of variation around the mean. For all practical purposes we recommend to center the data.

An alternative to mean-centering is to keep the origin (0-value for all variables) as model center. This is only advisable in the special case of a regression model where you would know in advance that the linear relationship between X and Y is supposed to go through the origin.

Note 1: Centering is ***included as a default option*** in the relevant analysis dialogs, and the computations are done as a ***first stage of the analysis***.

Note 2: Mean centering is also available as a ***transformation*** to be performed ***manually*** from the Editor. This allows you for instance to plot the centered data.

Weighting

PCA, PLS and PCR are projection methods based on finding directions of maximum variation. Thus, they all depend on the ***relative variance*** of the variables.

Depending on the kind of information you want to extract from your data, you may need to use weights based on the standard deviation of the variables, i.e. square root of variance, which expresses the variance in the same unit as the original variable. This operation is also called ***scaling***.

Note 1: Weighting is ***included as a default option*** in the relevant analysis dialogs, and the computations are done as a ***first stage of the analysis***.

Note 2: Standard deviation scaling is also available as a ***transformation*** to be performed ***manually*** from the Editor. This may help you study the data in various plots from the Editor, or prior to computing descriptive statistics. It may for example allow you to compare the distributions of variables of different scales into one plot.

Weighting Options in The Unscrambler

The following weighting options are available in the analysis dialogs of The Unscrambler:

- 1
- 1/Sdev
- Constant
- A/Sdev+B
- Passify

Weighting Option: 1

1 represents no weighting at all, i.e. all computations are based on the ***raw variables***.

Weighting Option: 1/SDev

1/SDev is called ***standardization*** and is used to give all variables the same variance, i.e. 1. This gives all variables the same chance to influence the estimation of the components, and is often used if the variables

- are measured with different units;
- have different ranges;
- are of different types.

Sensory data, which are already measured in the same units, are nevertheless sometimes standardized if the scales are used differently for different attributes.

Caution! If a noisy variable with small standard deviation is standardized, its influence will be increased, which can sometimes make the model less reliable.

Weighting Option: Constant

This option can be used to set the weighting for each variable manually.

Weighting Option: A/Sdev+B

A/SDev+B can be used as an alternative to full standardization when this is considered to be too dangerous. It is a compromise between 1/SDev and a constant.

Application:

To keep a noisy variable with a small standard deviation in an analysis while reducing the risk of "blowing up noise", use A/Sdev + B with a value of A smaller than 1, and / or a non-zero value of B.

Weighting Option: Passify

Projection methods (PCA, PCR and PLS) take advantage of variances and covariances to build models where the influence of a variable is determined by its variance, and the relationship between two variables may be summarized by their correlation.

While variance is sensitive to weighting, correlation is not. This provides us with a possibility of still studying the relationship between one variable and the others, while limiting this variable's influence on the model. This is achieved by giving this variable a ***very low weight*** in the analysis. This operation is called ***Passifying*** the variable.

Passified variables will lose any influence they might have on the model, but by plotting Correlation Loadings you will have a chance to study their behavior in relation to the active variables.

Weighting: The Case of PLS2 and PLS1

For PLS2, the X- and Y-matrices can be weighted independently of each other, since only the relative variances inside the X-matrix and the relative variances inside the Y-matrix influence the model.

Even if weighting of Y has no effect on a PLS1 model, it is useful to get X and Y in the same scale in the result plots.

Weighting: The Case of Sensory Analysis

There is disagreement in the literature about whether one should standardize sensory attributes or use them as they are. Generally, this decision depends on how the assessors are trained, and also on what kind of information the analysis is supposed to give.

A standardization corresponds to a stretching/shrinking that gives new "sensory scores" which measure position relative to the extremes in the actual data table. In other words, ***standardization*** of variables gives an analysis that interprets the variation ***relative to the extremes*** in the data table.

The opposite, ***no weighting*** at all, gives an analysis that has a closer relationship to the ***individual assessor's*** personal extremes, and these are strongly related to their very subjective experience and background.

We therefore generally recommend standardization. This procedure, however, has an important disadvantage: It may increase the relative influence of unreliable or noisy attributes (see *Caution* in section Weighting Option: 1/SDev).

Weighting: The Case of Spectroscopy Data

Standardization of spectra may make it more difficult to interpret loading plots, and you risk blowing up noise in wavelengths with little information. Thus, spectra are generally not weighted, but there are exceptions.

Weighting: The Case of Three-way Data

You will find special considerations about centering and weighting for three-way data in section Pre-processing of Three-way Data.

Pre-processing of Three-way Data

Pre-processing of three-way data requires some attention as shown by Bro & Smilde 2003 (see detailed bibliography given in the Method References chapter). The main objective of pre-processing is to simplify subsequent modelling. Certain types of centering and scaling in three-way analysis may lead to the opposite effect because they can introduce artificial variation in the data.

From a user perspective the differences from two-way pre-processing are not too problematic because The Unscrambler has been adapted to make sure that only proper pre-processing is possible.

Centering and Weighting for Three-way Data

Centering is performed to make the data compatible with the structural model (remove non-trilinear parts). Scaling (weighting) on the other hand is a way of making the data compatible with the least squares loss function normally used. Scaling does not change the structural model of the data, but only the weight paid to errors of specific elements in the estimation (see Bro 1998 - detailed bibliography given in the Method References chapter). Centering must be done across the columns of the matrix, i.e. a scalar is subtraced from each column. Scaling has to be done on the rows, that is, all elements of a row are divided by the same scalar.

The main issue in pre-processing of three-way arrays in regression models is that scaling should be applied on each mode separately. It is not useful or sensible to scale three-way data when it is rearranged into a matrix. In order to scale data to something similar to auto-scaling, standardization has to be imposed for both variable modes.

Re-formatting and Pre-processing in Practice

This chapter lists menu options and dialogs for data re-formatting and transformations. For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo's web site www.camo.com/TheUnscrambler/Appendices .

Make Simple Changes In The Editor

From the Editor, you can make changes to a data table in various ways, through two menus:

1. The **Edit** menu lets you move your data through the clipboard and modify your data table by inserting or deleting samples or variables.
2. The **Modify** menu includes two options which allow you to change variable properties.

Copy / Paste Operations

- **Edit - Cut**: Remove data from the table and store it on the clipboard
- **Edit - Copy**: Copy data from the table to the clipboard
- **Edit - Paste**: Paste data from the clipboard to the table

Add or Delete Samples / Variables

- **Edit - Insert - Sample**: Add new sample above cursor position

- **Edit - Insert - Variable**: Add new variable left to cursor position
- **Edit - Insert - Category Variable**: Add new category variable left to cursor position
- **Edit - Insert - Mixture Variables**: Add new mixture variables left to cursor position
- **Edit - Append - Samples**: Add new samples at the end of the table
- **Edit - Append - Variables**: Add new variable at the end of the table
- **Edit - Append - Category Variable**: Add new category variable at the end of the table
- **Edit - Append - Mixture Variables**: Add new mixture variables at the end of the table
- **Edit - Delete**: Delete selected sample(s) / variable (s)

Change Data Values

- **Edit - Fill…**: Fill selected cells with a value of your choice
- **Edit - Fill Missing**: Fill empty cells with values estimated from the structure in the non-missing data
- **Edit - Find/Replace…**: Find cells with requested value and replace

Operations on Category Variables

- **Edit - Convert to Category Variable**: Convert from continuous to category (discrete or ranges)
- **Edit - Split Category Variable**: Convert from category to indicator (binary) variables
- **Modify - Properties**: Change name and levels

Operations on Mixture Variables

- **Edit - Convert to Mixture Variable**: Convert from continuous to mixture
- **Edit - Correct Mixture Components**: Ensure that sum of mixture components is equal to "Mixsum" for each sample

Locate or Select Cells

- **Edit - Go To**: Go to desired cell
- **Edit - Select Samples…**: Select desired samples
- **Edit - Select Variables…**: Select desired variables
- **Edit - Select All**: Select the whole table contents

Display and Formatting Options

- **Edit - Adjust Width**: Adjust column width to displayed values
- **Modify - Properties**: Change name of selected sample or variable and lookup general properties
- **Modify - Layout…**: Change display format of selected variable

The Editor: The Case of 3-D Data Tables

3-D data tables are physically stored in an unfolded format, and displayed accordingly in the Editor. For instance, a 3-way array (4x5x2) with OV^2 layout will be stored as a matrix with 4 rows and 5x2=10 columns. In the Editor, it will appear as a 3-D table with 4 samples, 5 Primary variables and 2 Secondary variables.

This has the advantage of displaying all data values in one window. No need to look at several sheets to get a full overview!

Some existing features accessible from the Editor have been adapted to 3-D data, and specific features have been developed (see for instance section “Change the Layout or Order of Your Data” below).

However, some features which do not make sense for three-way data, or which would introduce inconsistencies in the 3-D structure, are not available when editing 3-D data tables. Lookup Chapter “Re-formatting and Pre-processing: Restrictions for 3D Data Tables” p.88 for an overview of those limitations.

Organize Your Samples And Variables Into Sets

The **Set Editor**, which enables you to define groups of variables or samples that belong together and to add interactions and squares to a group of variables, is available from the **Modify** menu.

- **Modify - Edit Set**: Define new sample or variable sets or change their definition

Change the Layout or Order of Your Data

Various options from the **Modify** menu allow you to change the order of samples or variables, as well as more drastically modifying the layout (2-D or 3-D) of your data table.

Sorting Operations

- **Modify - Sort Samples…**: Sort samples according to name or values of some variables
- **Modify - Sort Samples by Sets**: Group samples according to which set they belong
- **Modify - Sort Variables by Sets**: Group variables according to which set they belong
- **Modify - Reverse Sample Order**: Sort samples from last to first
- **Modify - Reverse Variable Order**: Sort variables from last to first

Change Table Layout

- **Modify - Transform - Transpose**: Samples become variables and variables become samples
- **Modify - Swap 3-D Layout**: Switch 3-D data from OV2 to O2V or vice-versa
- **Modify - Swap Samples & Variables**: 6 options for swapping samples and variables in a 3-D data table
- **Modify - Toggle 3-D Layouts**: Quick change of layout for a 3-D data table
- **File - Duplicate - As 2-D Data Table**: Unfold 3-D data to a 2-D structure
- **File - Duplicate - As 3-D Data Table**: Build a 3-D data table from an unfolded 2-D structure

Apply Transformations

Transform your samples or variables to make their properties more suitable for analysis and easier to interpret. Apply ready-to-use transformations or make your own computations.

Bilinear models, e.g. PCA and PLS, basically assume linear data. Therefore, if you have non-linearities in your data, you may apply transformations which result in a more symmetrical distribution of the data and a better fit to a linear model.

Note: Transformations which may change the dimensions of your data table are disabled for 3-D data tables.

General Transformations

- **Modify - Compute General**: Apply simple arithmetical or mathematical operations (+, *, log…)
- **Modify - Transform - Noise**: Add noise to your data so as to test model robustness

Transformations Based on Curves or Vectors

- **Modify - Shift Variables…**: Create time lags by shifting variables up or down
- **Modify - Transform - Smoothing**: Reduce noise by smoothing the curve formed by a series of variables
- **Modify - Transform - Normalize**: Scale the samples by applying normalization to a series of variables
- **Modify - Transform - Spectroscopic Transformation**: Change spectroscopic units
- **Modify - Transform - MSC/EMSC**: Remove scatter or baseline effects
- **Modify - Transform - Derivatives**: Compute derivatives of the curve formed by a series of variables
- **Modify - Transform - Baseline**: Baseline Correction for spectra
- **Modify - Transform - SNV**: Center and scale individual spectra with Standard Normal Variate
- **Modify - Transform - Center and Scale**: Apply mean centering and/or standard deviation scaling
- **Modify - Transform - Reduce (Average)**: Average over a number of adjacent samples or variables

User-defined Transformations

- **Modify - Transform - User-defined**: Apply a transformation programmed outside The Unscrambler

Undo and Redo

Many re-formatting or pre-processing operations done through the **Edit** and **Modify** menus can be ***undone*** or ***redone***.

- **Modify - Undo**: Undo the last editing operation
- **Modify - Redo**: Re-apply the undone operation

Re-formatting and Pre-processing: Restrictions for 3D Data Tables

The following operations are disabled in the case of 3-D data tables:

- Operations which change the ***number*** or ***order*** of the samples (O^2V layout) or variables (OV^2 layout);
- Operations which have to do with ***mixture*** variables, since experimental design is not implemented for three-way arrays;
- ***User-defined*** transformations.

The following menu options may be affected by these restrictions:

Edit - Paste	**Modify - Reduce (Average)**
Edit - Insert	**Modify - Transpose**
Edit - Append	**Modify - User-defined**
Edit - Delete	**Modify - Sort Samples**
Edit - Convert to Category Variable	**Modify - Sort Samples/Variables by Sets**

Edit - Convert to Mixture Variable

Modify - Shift Variables

Modify - Reverse Sample/Variable Order

Re-formatting and Pre-processing: Restrictions for Mixture and D-Optimal Designs

The options from the **Modify** menu which are accessible to operate modifications on mixture and D-optimal designed data tables are:

- on ***Response*** variables, all operations can be performed
- on ***Process*** variables, all non re-sizing transformations can be performed.

You can operate the **Sort Samples** and **Shift Variables** options on Mixture variables contained in a Non-Designed data table, but not in a Designed data table.

Describe One Variable At A Time

Get to know each of your variables individually with descriptive statistics.

Simple Methods for Univariate Data Analysis

Throughout this chapter, we will consider a data table with one row for each object (or individual, or sample), and one column for each descriptor (or measure, or variable). The rows will be referred to as ***samples***, and the columns as ***variables***.

The methods described in the sections that follow will help you ***get better acquainted with your data***, so as to answer such questions as:

- How many cells in my data table are ***empty*** (missing values)?
- What are the ***minimum*** and ***maximum*** values of variable "Yield"?
- Does variable "Viscosity" follow a ***normal distribution***?
- Are there any extreme / unlikely / impossible values for some variables (suggesting ***data entry errors***)?
- What is the ***shape of the relationship*** between variables "Yield" and "Impurity %"?
- Do all panelists use the sensory scale in the same way (minimum, maximum, mean, standard deviation)?
- Are there any visible differences in ***average Yield*** between ***three production lines***?

Descriptive Statistics

Descriptive statistics is a summary of the distribution of one or two variables at a time. It is not supposed to tell much about the structure of the data, but it is useful if you want to get a quick look at each separate variable before starting an analysis.

- ***One-way statistics*** - mean, standard deviation, variance, median, minimum, maximum, lower and upper quartile - can be used to spot any out-of-range value, or to detect abnormal spread or asymmetry. You should check this before proceeding with any further analysis, and look into the raw data if they suggest anything suspect. A transformation might also be useful.
- ***Two-way statistics*** - correlations - show how the variations of two different variables are linked in the data you are studying.

First Data Check

Prior to any other analysis, you may use a few simple statistical measures directly from the Editor to ***check your data***. These analyses can be computed either on samples or on variables and include number of missing values, minimum, maximum, mean and standard deviation.

Checking these statistics is useful if you want to detect out-of-range values or pick out variables and samples that have too many missing values to be reliably included in a model.

Descriptive Variable Analysis

After you have performed the initial, simple checks, it might also be useful to get better acquainted with your data by computing more extensive statistics on the variables.

One-way and two-way statistics can be computed on any subset of your data matrix, with or without ***grouping*** according to the values of a ***leveled*** variable.

- For ***non-designed*** data tables, this means that you can group the samples according to the levels of one or several ***category*** variables.
- For ***designed data***, in addition to optional grouping according to the levels of the ***design variables***, ***predefined*** groups such as “Design Samples” or “Center Samples” are automatically taken into account.

Plots For Descriptive Statistics

The descriptive statistics can be displayed as plots.

- ***Line plots*** show mean or standard deviation, or mean and standard deviation together;
- ***Box-plots*** show the percentiles (min, lower quartile, median, upper quartile, max).

In addition, you may graphically study the correlation between two variables by plotting them as a ***2D scatter plot***. If you turn on Plot Statistics, the value of the ***correlation coefficient*** will be displayed among other information.

Univariate Data Analysis in Practice

This section lists menu options, dialogs and plots for descriptive statistics. For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo’s web site www.camo.com/TheUnscrambler/Appendices .

Display Descriptive Statistics In The Editor

You may display simple ***descriptive statistics*** on some of your variables or samples directly from the ***Editor***. This is a quick way to check for instance how many values are missing or whether the maximum value of a variable is outside the expected range, indicating a probable error in the data.

- **View - Sample Statistics**: Display descriptive statistics for your samples in a slave Editor window
- **View - Variable Statistics**: Display descriptive statistics for your variables in a slave Editor window

Study Your Variables Graphically

Several types of ***plots*** of raw data, produced from the ***Editor***, allow you to get an overview of e.g. variable distributions, 2-variable correlation or sample spread.

Most Relevant Types of Plots

- **Plot - 2D Scatter**: Plot two variables (or samples) against each other
- **Plot - Normal Probability**: Plot one variable (or sample) and check against a normal distribution
- **Plot - Histogram**: Plot one variable (or sample) as number of elements in evenly spread ranges of values

Include More Information in your Plot

- **View - Plot Statistics**: Display useful statistics in 2D Scatter or Histogram plot
- **View - Trend Lines - Regression Line**: Add a regression line to your 2D Scatter Plot
- **View - Trend Lines - Target Line**: Add a target line to your 2D Scatter Plot

More About How To Use and Interpret Plots of Raw Data

Read about the following in chapter "Represent Data":

- Line Plot of Raw Data, p. **Feil! Bokmerke er ikke definert.**
- 2D Scatter Plot of Raw Data, p. 65
- 3D Scatter Plot of Raw Data, p. 65
- Matrix Plot of Raw Data, p. 66
- Normal Probability Plot of Raw Data, p. 66
- Histogram of Raw Data, p. 67

Compute And Plot Detailed Descriptive Statistics

When your data table is displayed in the ***Editor***, you may access the **Task** menu to run a suitable analysis. It is recommended to start with Descriptive Statistics before running more complex analyses.

Once the descriptive statistics have been computed according to your specifications, View the results and display them as plots from the ***Viewer***.

Details:

- **Task - Statistics**: Run the computation of Descriptive Statistics on a selection of variables and samples
- **Plot - Statistics**: Specify how to plot the results in the Viewer
- **Results - Statistics**: Retrieve Statistics results and display them in the Viewer

Describe Many Variables Together

Principal Component Analysis (PCA) summarizes the structure in large amounts of data. It shows you how variables co-vary and how samples differ from each other.

Principles of Descriptive Multivariate Analysis (PCA)

The purpose of descriptive multivariate analysis is to get the best possible view of the structure, i.e. the variation that makes sense, in the data table you are analyzing. PCA (***Principal Component Analysis***) is the method of choice.

Throughout this chapter, we will consider a data table with one row for each object (or individual, or sample), and one column for each descriptor (or measure, or variable). The rows will be referred to as ***samples***, and the columns as ***variables***.

Purposes Of PCA

Large data tables usually contain a large amount of information, which is partly hidden because the data are too complex to be easily interpreted. ***Principal Component Analysis*** (PCA) is a projection method that helps you visualize all the information contained in a data table.

PCA helps you find out in what respect one sample is different from another, which variables contribute most to this difference, and whether those variables contribute in the same way (i.e. are correlated) or independently from each other. It also enables you to detect sample patterns, like any particular grouping.

Finally, it quantifies the amount of ***useful information*** - as opposed to noise or meaningless variation - contained in the data.

It is important that you understand PCA, since it is a very useful method in itself, and forms the basis for several classification (SIMCA) and regression (PLS/PCR) methods. The following is a brief introduction; we refer you to the book “Multivariate Analysis in Practice” by Kim Esbensen *et al.*, and other references given in the Method References chapter for further reading.

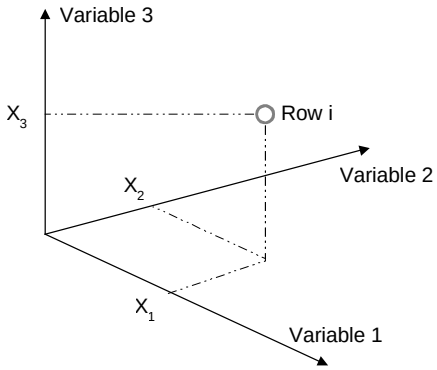
How PCA Works (In Short)

To understand how PCA works, you have to remember that ***information*** can be assimilated to ***variation***. Extracting information from a data table means finding out what makes one sample different from - or similar to - another.

Geometrical Interpretation Of Difference Between Samples

Let us look at each sample as a point in a multidimensional space (see figure below). The location of the point is determined by its coordinates, which are the cell values of the corresponding row in the table. Each variable thus plays the role of a coordinate axis in the multidimensional space.

The sample in multidimensional space



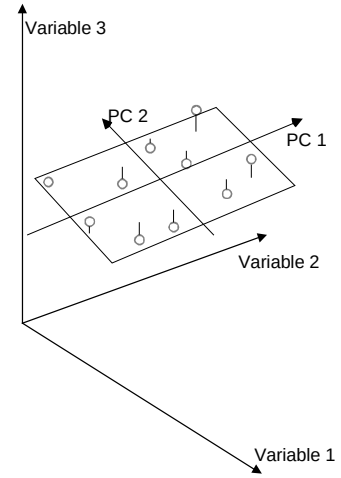
Let us consider the whole data table geometrically. Two samples can be described as ***similar*** if they have close values for most variables, which means close coordinates in the multidimensional space, i.e. the two points are located in the same area. On the other hand, two samples can be described as ***different*** if their values differ a lot for at least some of the variables, i.e. the two points have very different coordinates, and are located far away from each other in the multidimensional space.

Principles Of Projection

Bearing that in mind, the principle of PCA is the following: Find the directions in space along which the distance between data points is the largest. This can be translated as finding the linear combinations of the initial variables that contribute most to making the samples different from each other.

These directions, or combinations, are called ***Principal Components*** (PCs). They are computed iteratively, in such a way that the first PC is the one that carries most information (or in statistical terms: most explained variance). The second PC will then carry the maximum share of the residual information (i.e. not taken into account by the previous PC), and so on.

PCs 1 and 2 in a multidimensional space



This process can go on until as many PCs have been computed as there are variables in the data table. At that point, all the variation between samples has been accounted for, and the PCs form ***a new set of coordinate axes*** which has two advantages over the original set of axes (the original variables). First, the PCs are ***orthogonal*** to each other (we will not try to prove this here). Second, they are ***ranked*** so that each one carries more information than any of the following ones. Thus, you can prioritize their interpretation: Start with the first ones, since you know they carry more information!

The way it was generated ensures that this new set of coordinate axes is the most suitable basis for a graphical representation of the data that allows easy interpretation of the data structure.

Separating Information From Noise

Usually, only the ***first PCs*** contain genuine ***information***, while the later PCs most likely describe ***noise***. Therefore, it is useful to study the first PCs only instead of the whole raw data table: not only is it less complex, but it also ensures that noise is not mistaken for information.

Validation is a useful tool to make sure that you retain only informative PCs (see Chapter Principles of Model Validation p. 121 for details).

Is PCA the Most Relevant Summary of Your Data?

PCA produces an ***orthogonal*** bilinear matrix decomposition, where components or factors are obtained in a sequential way explaining maximum variance. Using these constraints plus normalization during the bilinear matrix decomposition, PCA produces ***unique*** solutions. These 'abstract' unique and orthogonal (independent) solutions are very helpful in deducing the number of different ***sources of variation*** present in the data and, eventually, they allow for their identification and interpretation. However, these solutions are 'abstract' solutions in the sense that they are not the 'true' underlying factors causing the data variation, but orthogonal linear combinations of them.

In some cases you might be interested in finding the ***'true' underlying sources of data variation***. It is not only a question of ***how many*** different sources are present and how they can be interpreted, but to find out ***how they are*** in reality. This can be achieved using another type of bilinear method called Curve Resolution. The price to pay is that Curve Resolution methods usually do not yield a unique solution unless external information is provided during the matrix decomposition.

Read more about Curve Resolution methods in the Help chapter “Multivariate Curve Resolution” p. 161.

Calibration, Validation and Related Samples

Any multivariate analysis - including PCA, and also regression - should include some validation (i.e. testing) to make sure that its results can be extrapolated to new data. This requires two separate steps in the computation of each model component (PC):

1. ***Calibration***: Finding the new component;
2. Validation: Checking whether the component describes new data well enough.

Each of those two steps requires its own set of samples; thus, we will later refer to ***calibration samples*** (or training samples), and to ***validation samples*** (or test samples).

A more detailed description of validation techniques and their interpretation is to be found in Chapter Validate A Model p. 121.

Main Results Of PCA

Each component of a PCA model is characterized by three complementary sets of attributes:

- ***Variances*** are error measures; they tell you how much information is taken into account by the successive PCs;
- ***Loadings*** describe the relationships between variables;
- ***Scores*** describe the properties of the samples.

Variances

The importance of a principal component is expressed in terms of variance. There are two ways to look at it:

- ***Residual variance*** expresses how much variation in the data ***remains to be explained*** once the current PC has been taken into account.
- ***Explained variance***, often measured as a percentage of the total variance in the data, is a measurement of the proportion of variation in the data ***accounted for*** by the current PC.

These two points of view are complementary. The variance which is not explained is residual.

These variances can be considered either for a single variable or sample, or for the whole data. They are computed as a mean square variation, with a correction for the remaining degrees of freedom.

Variances tell you how much of the information in the data table is being described by the model. The way they vary according to the number of model components can be studied to decide how complex the model should be (see section How To Use Residual And Explained Variances for more details).

Loadings

Loadings describe the data structure in terms of ***variable correlations***.

Each variable has a loading on each PC. It reflects both how much the variable contributed to that PC, and how well that PC takes into account the variation of that variable over the data points.

In geometrical terms, a loading is the cosine of the angle between the variable and the current PC: the smaller the angle (i.e. the higher the link between variable and PC), the larger the loading. It also follows that loadings can range between –1 and +1.

The basic principles of interpretation are the following:

1. For each PC, look for variables with ***high loadings*** (i.e. close to +1 or –1); this tells you the meaning of that particular PC (useful for further interpretation of the sample scores).
2. To study ***variable correlations***, use their loadings to imagine what their angles would look like in the multidimensional space. For instance, if two variables have high loadings along the same PC, it means that their angle is small, which in turn means that the two variables are highly correlated. If both loadings have the same sign, the correlation is positive (when one variable increases, so does the other). Else, it is negative (when one variable increases, the other decreases).

For more information on score and loading interpretation, see section How To Interpret PCA Scores And Loadings p.102, and examples in Tutorial B.

Scores

Scores describe the data structure in terms of sample patterns, and more generally show ***sample differences or similarities***.

Each sample has a score on each PC. It reflects the sample location along that PC; it is the coordinate of the sample on the PC.

You can interpret scores as follows:

1. Once the information carried by a PC has been interpreted with the help of the loadings, the score of a sample along that PC can be used to characterize that sample. It describes the ***major features*** of the sample, relative to the variables with high loadings on the same PC;

2. Samples with ***close scores*** along the same PC are ***similar*** (they have close values for the corresponding variables). Conversely, samples for which the scores differ much are quite different from each other with respect to those variables.

For more information on score and loading interpretation, see section How To Interpret PCA Scores And Loadings p.102, and examples in Tutorial B.

More Details About The Theory Of PCA

Let us have a more thorough look at PCA modeling to understand how you can diagnose and refine your PCA model.

The PCA Model As Approximation Of Reality

The underlying idea in PCA modeling is to replace a complex multidimensional data set by a simpler version involving ***fewer dimensions***, but still fitting the original data closely enough to be considered a ***good approximation***.

If you chose to retain all PCs, there would be no approximation at all - but then there would not be any gain in simplicity either! So deciding on the number of components to retain in a PCA model is **a trade-off between simplicity and completeness**.

Structure vs. Error

In matrix representation, the model with a given number of components has the following equation:

$$X = TP^T + E$$

where T is the scores matrix, P the loadings matrix and E the error matrix.

The combination of scores and loadings is the ***structure part*** of the data, the part that makes sense. What remains is called ***error*** or residual, and represents the fraction of variation that cannot be interpreted.

When you interpret the results of a PCA, you focus on the structure part and discard the residual part. It is OK to do so, provided that the residuals are indeed negligible. You decide yourself how large an error you can accept.

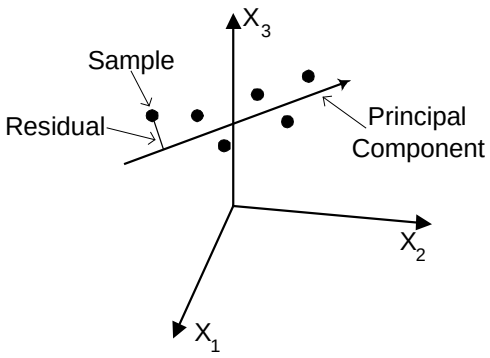
Sample Residuals

If you look at your data from the samples' point of view, each data point is approximated by another point which lies on the hyperplane generated by the model components.

The difference between the original location of the point and its approximated location (or ***projection*** onto the model) is the ***sample residual*** (see figure below).

This overall residual is a vector that can be decomposed in as many numbers as there are components. Those numbers are the sample residuals for each particular component.

Sample residuals



Variable Residuals

From the variables' point of view, the original variable vectors are being approximated by their projections onto the model components. The difference between the original vector and the projected one is the ***variable residual***.

It can also be broken down into as many numbers as there are components.

Residual Variation

The ***residual variation*** of a sample is the sum of squares of its residuals for all model components. It is geometrically interpretable as the squared distance between the original location of the sample and its projection onto the model.

The residual variations of Variables are computed the same way.

Residual Variance

The ***residual variance*** of a variable is the mean square of its residuals for all model components. It differs from the residual variation by a factor which takes into account the remaining degrees of freedom in the data, thus making it a valid expression of the ***modeling error*** for that variable.

Total residual variance is the average residual variance over all variables. This expression summarizes the overall modeling error, i.e. it is the variance of the error part of the data.

Explained Variance

Explained variance is the complement of residual variance, expressed as a percentage of the global variance in the data. Thus the explained variance of a variable is the fraction of the global variance of the variable taken into account by the model.

Total explained variance measures how much of the original variation in the data is described by the model. It expresses the ***proportion of structure*** found in the data by the model.

How To Interpret PCA Results

Once a model is built, you have to ***diagnose*** it, i.e. assess its quality, before you can actually use it for ***interpretation***.

There are two major steps in diagnosing a PCA model:

1. Check ***variances***, to determine how many components the model should include and know how much information the selected components take into account. At that stage, it is especially important to check validation variances (see Chapter Principles of Model Validation p. 121 for details on validation methods).
2. Look for ***outliers***, i.e. samples that do not fit into the general pattern.

These two steps may have to be run several times before you are satisfied with your model.

How To Use Residual And Explained Variances

Total Variances

Total residual and explained variances show how well the model ***fits*** to the data.

Models with small total residual variance (close to 0) or large total explained variance (close to 100%) explain most of the variation in the data. Ideally, you would want to have simple models, i.e. models where the residual variance goes down to zero with as few components as possible. If this is not the case, it means that there may be a large amount of noise in your data or, alternatively, that the data structure may be too complex to be accounted for by only a small number of components.

Variable Variances

Variables with ***small residual variance*** (or large explained variance) for a particular component are ***well explained*** by the corresponding model. Variables with large residual variance for all or for the 3-4 first components have a small or moderate relationship with the other variables.

If some variables have much larger residual variance than the other variables for all components (or for the first 3-4 of them), try to keep these variables out and make a new calculation. This may produce a model which is easier to interpret.

Calibration vs. Validation Variance

The calibration variance is based on fitting the calibration data to the model. The validation variance is computed by testing the model on data not used in building the model. Look at both variances to evaluate their difference. If the difference is large, there is reason to question whether the calibration data or the test data are representative.

Outliers can sometimes be the reason for large residual variance. The next section tells you more about outliers.

How To Detect Outliers In PCA

An ***outlier*** is a sample which looks so different from the others that it either is not well described by the model or influences the model too much. As a consequence, it is possible that one or more of the model components focus only on trying to describe how this sample is different from the others, even if this is irrelevant to the more important structure present in the other samples.

In PCA, outliers can be detected using score plots, residuals and leverages.

Different types of outliers can be detected by each tool:

- ***Score plots*** show sample patterns according to one or two components. It is easy to spot a sample lying far away from the others. Such samples are likely to be outliers.
- ***Residuals*** measure how well samples or variables fit the model determined by the components. Samples with a high residual are poorly described by the model, which nevertheless fits the other samples quite well. Such samples are strangers to the family of samples well described by the model, i.e. outliers.

- ***Leverages*** measure the distance from the projected sample (i.e. its model approximation) to the center (mean point). Samples with high leverages have a stronger influence on the model than other samples; they may or may not be outliers, but they are influential. An **influential outlier** (high residual + high leverage) is the worst case; it can however easily be detected using an **influence plot**.

How To Interpret PCA Scores And Loadings

Loadings show how data values vary when you move along a model component. This interpretation of a PC is then used to understand the meaning of the scores.

To figure out how this works, you must remember that the PCs are oriented axes. Loadings can have negative or positive values; so can scores. PCs build a link between samples and variables by means of scores and loadings.

First, let us consider one PC at a time. Here are the rules to interpret that link:

- If a variable has a ***very small loading***, whatever the sign of that loading, you should not use it for interpretation, because that variable is badly accounted for by the PC. Just discard it and focus on the variables with large loadings;
- If a variable has a ***positive loading***, it means that all samples with positive scores have higher than average values for that variable. All samples with negative scores have lower than average values for that variable;
- If a variable has a ***negative loading***, it means just the opposite. All samples with positive scores have lower than average values for that variable. All samples with negative scores have higher than average values for that variable;
- The higher the positive score of a sample, the larger its values for variables with positive loadings and vice versa;
- The more negative the score of a sample, the smaller its values for variables with positive loadings and vice versa;
- The larger the loading of a variable, the quicker sample values will increase with their scores.

To summarize, if the score of a sample and the loading of a variable on a particular PC have the same sign, the sample has higher than average value for that variable and vice-versa. The larger the scores and loadings, the stronger that relation.

If you now consider ***two PCs simultaneously***, you can build a 2-vector loading plot and a 2-vector score plot. The same principles apply to their interpretation, with a further advantage: you can now interpret any direction in the plot - not only the principal directions.

PCA in Practice

In practice, building and using a PCA model involves 3 steps:

1. Choose and implement an appropriate pre-processing method (see Chapter Re-formatting and Pre-processing p. 71);
2. Run the PCA algorithm, choose the number of components, diagnose the model;
3. Interpret the loadings and scores plots.

The sections that follow list menu options and dialogs for data analysis and result interpretation using PCA. For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo's web site www.camo.com/TheUnscrambler/Appendices.

Run A PCA

When your data table is displayed in the ***Editor***, you may access the **Task** menu to run a suitable analysis – for instance, PCA.

- **Task - PCA**: Run a PCA on the current data table

Save And Retrieve PCA Results

Once the PCA has been computed according to your specifications, you may either View the results right away, or Close (and Save) your PCA result file to be opened later in the ***Viewer***.

Save Result File from the Viewer

- **File - Save**: Save result file for the first time, or with existing name
- **File - Save As**: Save result file under a new name

Open Result File into a new Viewer

- **File - Open**: Open any file or just lookup file information
- **Results - PCA**: Open PCA result file or just lookup file information, warnings and variances
- **Results - All**: Open any result file or just lookup file information, warnings and variances

View PCA Results

Display PCA results as plots from the ***Viewer***. Your PCA results file should be opened in the Viewer; you may then access the **Plot** menu to select the various results you want to plot and interpret.

From the **View**, **Edit** and **Window** menus you may use more options to enhance your plots and ease result interpretation.

How To Plot PCA Results

- **Plot - PCA Overview**: Display the 4 main PCA plots
- **Plot - Variances and RMSEP**: Plot variance curves
- **Plot - Sample Outliers**: Display 4 plots for diagnosing outliers
- **Plot - Scores and Loadings**: Display scores and loadings separately or as a bi-plot
- **Plot - Scores**: Plot scores along selected PCs
- **Plot - Loadings**: Plot loadings along selected PCs
- **Plot - Residuals**: Display various types of residual plots
- **Plot - Leverage**: Plot sample leverages

How To Display Uncertainty Results

- **View - Hotelling T2 Ellipse**: Display Hotelling T^2 ellipse on a score plot
- **View - Uncertainty Test - Stability Plot**: Display stability plot for scores or loadings
- **View - Correlation Loadings**: Change a loading plot to display correlation loadings

PC Navigation Tool

Navigate up or down the PCs in your model along the vertical and horizontal axes of your plots:

- **View - Source - Previous Vertical PC**
- **View - Source - Next Vertical PC**
- **View - Source - Back to Suggested PC**
- **View - Source - Previous Horizontal PC**
- **View - Source - Next Horizontal PC**

More Plotting Options

- **View - Source**: Select which sample types / variable types / variance type to display
- **Edit - Options**: Format your plot
- **Edit - Insert Draw Item**: Draw a line or add text to your plot
- **View - Outlier List**: Display list of outlier warnings issued during the analysis for each PC, sample and/or variable
- **Window - Warning List**: Display general warnings issued during the analysis
- **View - Toolbars**: Select which groups of tools to display on the toolbar
- **Window - Identification**: Display curve information for the current plot

How To Change Plot Ranges:

- **View - Scaling**
- **View - Zoom In**
- **View - Zoom Out**

How To Keep Track of Interesting Objects

- **Edit - Mark**: Several options for marking samples or variables

How To Display Raw Data

- **View - Raw Data**: Display the source data for the analysis in a slave Editor

Run New Analyses From The Viewer

In the ***Viewer***, you may not only **Plot** your PCA results; the **Edit - Mark** menu allows you to mark samples or variables that you want to keep track of (they will then appear marked on all plots), while the **Task - Recalculate…** options make it possible to re-specify your analysis without leaving the viewer.

Check that the currently active subview contains the right type of plot (samples or variables) before using **Edit - Mark**.

How To Keep Track of Interesting Objects

- **Edit - Mark - One By One**: Mark samples or variables individually on current plot
- **Edit - Mark - With Rectangle**: Mark samples or variables by enclosing them in a rectangular frame (on current plot)
- **Edit - Mark - Outliers Only**: Mark automatically detected outliers
- **Edit - Mark - Test Samples Only**: Mark test samples (only available if you used test set validation)
- **Edit - Mark - Evenly Distributed Samples Only**: Mark a subset of samples which evenly cover your data range

How To Remove Marking

- **Edit - Mark - Unmark All**: Remove marking for all objects of the type displayed on current plot

How To Reverse Marking

- **Edit - Mark - Reverse Marking**: Exchange marked and unmarked objects on the plot

How To Re-specify your Analysis

- **Task - Recalculate with Marked**: Recalculate model with only the marked samples / variables
- **Task - Recalculate without Marked**: Recalculate model without the marked samples / variables
- **Task - Recalculate with Passified Marked**: Recalculate model with marked variables weighted down using “Passify”
- **Task - Recalculate with Passified Unmarked**: Recalculate model with unmarked variables weighted down using “Passify”

Extract Data From The Viewer

From the ***Viewer***, use the **Edit - Mark** menu to mark samples or variables that you have reason to single out, e.g “dominant variables” or “outlying samples”, etc.

There are two ways to display the source data for the currently viewed analysis into a new Editor window.

1. Command **View - Raw Data** displays the source data into a ***slave*** Editor table, which means that marked objects on the plots result in highlighted rows (for marked samples) or columns (variables) in the Editor. If you change the marking, the highlighting will be updated; if you highlight different rows or columns, you will see them marked on the plots.
2. You may also take advantage of the **Task - Extract Data…** options to display raw data for only the samples and variables you are interested in. A new data table is created and displayed in an independent Editor window. You may then edit or re-format those data as you wish.

How To Mark Objects

- Lookup the previous section **View - Raw Data**: Display the source data for the analysis in a slave Editor

Run New Analyses From The Viewer.

How To Display Raw Data

- **View - Raw Data**: Display the source data for the analysis in a slave Editor

How To Extract Raw Data

- **Task - Extract Data from Marked**: Extract data for only the marked samples / variables
- **Task - Extract Data from Unmarked**: Extract data for only the unmarked samples / variables

How to Run an Analysis on 3-D Data

PCA is disabled for 3-D data; however, three-way PLS (or tri-PLS) is available as a three-way regression method. Look it up in Chapter Three-way Data Analysis.

Useful tips

To run a PCA on your 3-way data, you need to ***duplicate your 3-D table as 2-D data*** first. Then all relevant analyses will be enabled.

For instance, you may run a PCA on unfolded 3-way spectral data, by doing the following sequence of operations:

1. Start from your 3-D data table (OV^2 layout) where each row contains a 2-way spectrum;
2. Use **File - Duplicate - As 2-D Data Table**: this generates a 2-D table containing unfolded spectra;
3. Save the resulting 2-D table with **File - Save As**;
4. Use **Task - PCA** to run the desired analysis.

Another possibility is to develop your own three-way analysis routine and implement it as a User-Defined Analysis (UDA). Such analyses may then be run from the **Task - User-defined Analysis** menu.

Combine Predictors and Responses In A Regression Model

Principles of Predictive Multivariate Analysis (Regression)

Find out about how well some predictor variables (X) explain the variations in some response variables (Y) using MLR, PCR, PLS, or nPLS.

Note: The sections in this chapter focus on methods dealing with two-dimensional data stored in a 2-D data table.

If you are interested in ***three-way modeling***, adapted to three-way arrays stored in a 3-D data table, you may first read this chapter so as to learn about the general principles of regression, then go to Chapter "Three-way Data Analysis" p. 177 where these principles will be taken further so as to apply to your case.

What Is Regression?

Regression is a generic term for all methods attempting to fit a model to observed data in order to ***quantify the relationship*** between two groups of variables. The fitted model may then be used either to merely ***describe*** the relationship between the two groups of variables, or to ***predict*** new values.

General Notation and Definitions

The two data matrices involved in regression are usually denoted X and Y, and the purpose of regression is to build a model ***Y = f(X)***. Such a model tries to explain, or predict, the variations in the Y-variable(s) from the variations in the X-variable(s). The link between X and Y is achieved through a common set of samples for which both X- and Y-values have been collected.

Names for X and Y

The X- and Y-variables can be denoted with a variety of terms, according to the particular context (or culture). most common ones:

Usual names for X- and Y-variables

Context	X	Y
General	Predictors	Responses
Multiple Linear Regression (MLR)	Independent Variables	Dependent Variables
Designed Data	Factors, Design Variables	Responses
Spectroscopy	Spectra	Constituents

Univariate vs. Multivariate Regression

Univariate regression uses a single predictor, which is often not sufficient to model a property precisely. ***Multivariate regression*** takes into account several predictive variables simultaneously, thus modeling the property of interest with more accuracy.

The whole chapter focuses on multivariate regression.

How And Why To Use Regression

Building a regression model involves collecting predictor and the response values for common samples, and then fitting a predefined mathematical relationship to the collected data.

For example, in analytical chemistry, spectroscopic measurements are made on solutions with known concentrations of a given compound. Regression is then used to relate ***concentration*** to ***spectrum***.

Once you have built a regression model, you can ***predict the unknown concentration*** for new samples, using the spectroscopic measurements as predictors. The advantage is obvious if the concentration is difficult or expensive to measure directly.

More generally, classical indications for regression as a predictive tool could be the following:

- Every time you wish to use cheap, easy-to-perform measurements as a ***substitute*** for more expensive or time-consuming ones;
- When you want to build a ***response surface model*** from the results of some experimental design, i.e. describe precisely the response levels according to the values of a few controlled factors.

What Is A Good Regression Model?

The purpose of a regression model is to extract ***all*** the ***information relevant for the prediction*** of the response from the available data.

Unfortunately, observed data usually contain some amount of noise, and may also include some irrelevant information:

- ***Noise*** can be random variation in the response due to experimental error, or it can be random variation in the data values due to measurement error. It may also be some amount of response variation due to factors that are not included in the model.
- ***Irrelevant information*** is carried by predictors that have little or nothing to do with the modeled phenomenon. For instance, NIR absorbance spectra may carry some information relative to the solvent and not only to the compound of which you are trying to predict the concentration.

A good regression model should be able to

- Pick up only relevant information, and all of it. It should leave aside irrelevant variation and focus on the fraction of variation in the predictors which affects the response;
- Avoid overfitting, i.e. distinguish between variation in the response that can be explained by variation in the predictors, and variation caused by mere noise.

Regression Methods In The Unscrambler

The Unscrambler contains three regression methods:

1. Multiple Linear Regression (MLR)
2. Principal Component Regression (PCR)
3. PLS Regression

Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is a well-known statistical method based on ordinary least squares regression. It estimates the model coefficients by the equation:

$$b = (X^T X)^{-1} X^T y$$

This operation involves a matrix inversion, which leads to ***collinearity*** problems if the variables are not linearly independent. Incidentally, this is the reason why the predictors are called independent variables in MLR; the ability to vary independently of each other is a crucial requirement to variables used as predictors with this method. MLR also requires more samples than predictors or the matrix cannot be inverted.

The Unscrambler uses Singular Value Decomposition to find the MLR solution. No missing values are accepted.

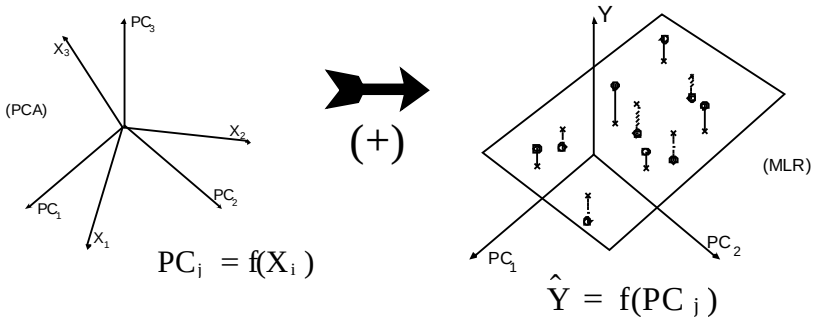
More About:

- How MLR compares to other regression methods in "More Details About Regression Methods" p.114
- MLR results in "Main Results Of Regression" p.111

Principal Component Regression (PCR)

Principal Component Regression (PCR) is a two-step procedure that first decomposes the X-matrix by PCA, then fits an MLR model, using the PCs instead of the original X-variables as predictors.

PCR procedure



More About:

- How PCR compares to other regression methods in "More Details About Regression Methods" p.114
- PCR results in "Main Results Of Regression" p.111

References:

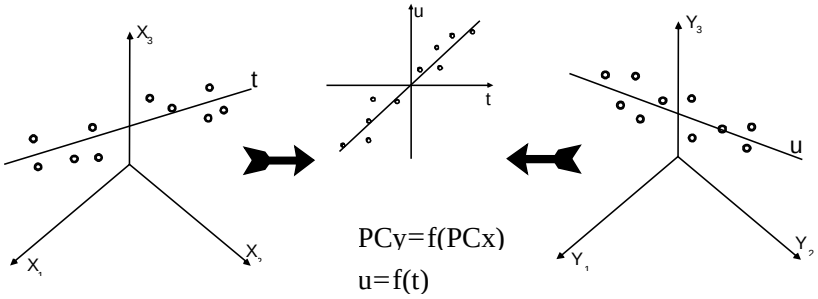
- Principles of Projection and PCA p. 95

You may also read about the PCR algorithm in the Method Reference chapter, available as a separate .PDF document for easy print-out of the algorithms and formulas – download it from Camo's web site www.camo.com/TheUnscrambler/Appendices.

PLS Regression

Partial Least Squares - or Projection to Latent Structures - (PLS) models both the X- and Y-matrices simultaneously to find the latent variables in X that will best predict the latent variables in Y. These PLS-components are similar to principal components, and will also be referred to as PCs.

PLS procedure



There are two versions of the PLS algorithm:

- ***PLS1*** deals with only one response variable at a time (like MLR and PCR);
- ***PLS2*** handles several responses simultaneously.

More About:

- How PLS compares to other regression methods in “More Details About Regression Methods” p.114
- PLS results in “Main Results Of Regression” p. 111

References:

- Principles of Projection and PCA p. 95

You may also read about the PLS1 and PLS2 algorithms in the Method Reference chapter, available as a separate .PDF document for easy print-out of the algorithms and formulas formulas – download it from Camo’s web site www.camo.com/TheUnscrambler/Appendices.

Calibration, Validation and Related Samples

All regression modeling should include some validation (i.e. testing) to make sure that its results can be extrapolated to new data. This requires two separate steps in the computation of each model component (PC):

1. ***Calibration***: Finding the new component;
2. ***Validation***: Checking whether the component describes new data well enough.

Calibration is the fitting stage in the regression modeling process: The main data set, containing only the calibration sample set, is used to compute the model parameters (PCs, regression coefficients).

We ***validate*** our models to get an idea of how well a regression model would perform if it were used to predict new, unknown samples. A test set consisting of samples with known response values is usually used. Only the X-values are fed into the model, from which response values are predicted and compared to the known, true response values. The model is validated if the prediction residuals are low.

Each of those two steps requires its own set of samples; thus, we will later refer to ***calibration samples*** (or **training** samples), and to ***validation samples*** (or **test** samples).

A more detailed description of validation techniques and their interpretation is to be found in Chapter "Validate A Model" p. 121.

Main Results Of Regression

The main results of a regression analysis vary depending on the method used. They may be roughly divided into two categories:

1. ***Diagnosis***: results that help you check the validity and quality of the model;
2. ***Interpretation***: results that give you insight into the shape of the relationship between X and Y, as well as (for projection methods only) sample properties.

Note that some results, e.g. scores, may be considered as belonging to both categories (scores can help you detect outliers, but they also give you information about differences or similarities among samples).

The table below lists the various types of regression results computed in The Unscrambler, their application area (diagnosis "D" or interpretation "I") and the regression method(s) for which they are available.

Regression results available for each method

Result	Application	MLR	PCR	PLS
B-coefficients	I	X	X	X
Predicted Y-values	I,D	X	X	X
Residuals (*)	D	X	X	X
Error Measures (*)	D	X	X	X
ANOVA	D	X		
Scores and Loadings (**)	I,D		X	X
Loading weights	I,D			X

(*) The various residuals and error measures are available for each PC in PCR and PLS, while for MLR there is only one of each type

(**) There are two types of scores and loadings in PLS, only one in PCR

In short, all three regression methods give you a model with an **equation** expressed by the regression coefficients (***b-coefficients***), from which ***predicted Y-values*** are computed. For all methods, ***residuals*** can be computed as the difference between predicted (fitted) values and actual (observed) values; these residuals can then be combined into ***error measures*** that tell you how well your model performs.

PCR and PLS, in addition to those standard results, provide you with powerful interpretation and diagnostic tools linked to **projection**: more elaborate error measures, as well as ***scores and loadings***.

The simplicity of MLR, on the other hand, allows for simple **significance testing** of the model with ***ANOVA*** and of the b-coefficients with a Student's test (ANOVA will not be presented hereafter; read more about it in the ANOVA section from Chapter "Analyze Results from Designed Experiments" p. 149.)

However, significance testing is also possible in PCR and PLS, using Martens' Uncertainty Test.

B-coefficients

The regression model can be written

$$Y = b_0 + b_1X_1 + ... + b_kX_k + e$$

meaning that the observed response values are approximated by a linear combination of the values of the predictors. The coefficients of that combination are called *regression coefficients* or ***B-coefficients***.

Several diagnostic tools are associated with the regression coefficients (available only for MLR):

- ***Standard error*** is a measure of the precision of the estimation of a coefficient;
- From then on, a Student's ***t-value*** can be computed;
- Comparing the t-value to a reference t-distribution will then yield a significance level or ***p-value***. It shows the probability of a t-value equal to or larger than the observed one would be if the true value of the regression coefficient were 0.

Predicted Y-values

Predicted Y-values are computed for each sample by applying the model equation with the estimated B-coefficients to the observed X-values.

For PCR or PLS models, the Predicted Y-values can also be computed using projection along the successive components of the model. This has the advantage of diagnosing samples which are badly represented by the model, and therefore have high prediction uncertainty. We will come back to this in Chapter "Make Predictions" p. 133.

Residuals

For each sample, the ***residual*** is the difference between observed Y-value and predicted Y-value. It appears as *e* in the model equation.

More generally, residuals may also be computed for each fitting operation in a projection model: thus the samples have X- and Y-residuals along each PC in PCR and PLS models. Read more about how sample and variable residuals are computed in Chapter "More Details About The Theory Of PCA" p. 99.

Error Measures for MLR

In MLR, all the X-variables are supposed to participate in the model independently of each other. Their co-variations are not taken into account, so X-variance is not meaningful there. Thus the only relevant measure of how well the model performs is provided by the *Y-variances*.

- ***Residual Y-variance*** is the variance of the Y-residuals and expresses how much variation remains in the observed response if you take out the modeled part. It is an overall measure of the misfit (i.e. the error made when you compute the fitted Y-value as a function of the X-values). It takes into account the remaining number of degrees of freedom in the data.
- ***Explained Y-variance*** is the complement to residual Y-variance, and is expressed as a percentage of the total Y-variance.
- ***RMSEC*** and ***RMSEP*** measure the calibration error and prediction error in the same units as the original response variable.

Residual and explained Y-variance are available for both calibration and validation.

Error Measures for PCR and PLS

In PCR and PLS models, not only the Y-variables are projected (fitted) onto the model; X-variables too! As mentioned previously, sample residuals are computed for each PC of the model. The residuals may then be combined

1. ***Across samples*** for each variable, to obtain a variance curve describing how the residual (or explained) variance of an **individual variable** evolves with the number of PCs in the model;

2. ***Across variables*** (all X-variables or all Y-variables), to obtain a Total variance curve describing the global fit of the model. The **Total Y-variance** curve shows how the prediction of Y improves when you add more PCs to the model; the **Total X-variance** curve expresses how much of the variation in the X-variables is taken into account to predict variation in Y.

Read more about how sample and variable residuals, as well as explained and residual variances, are computed in Chapter "More Details About The Theory Of PCA" p. 99.

In addition, the Y-***calibration error*** can be expressed in the same units as the original response variable using **RMSEC**, and the Y-***prediction error*** as **RMSEP**.

RMSEC and RMSEP also vary as a function of the number of PCs in the model.

Scores and Loadings (in General)

In PCR and PLS models, scores and loadings express how the samples and variables are projected along the model components.

PCR uses the same scores and loadings as PCA, since PCA is used in the decomposition of X. Y is then projected onto the "plane" defined by the MLR equation, and no extra scores or loadings are required to express this operation.

Read more about PCA scores and loadings in Chapters "Main Results Of PCA" p. 97 and "How To Interpret PCA Scores And Loadings" p. 102.

PLS scores and loadings are presented in the next two sections.

PLS Scores

Basically, PLS scores are interpreted the same way as PCA scores: They are the sample coordinates along the model components. The only new feature in PLS is that two different sets of components can be considered, depending on whether one is interested in summarizing the variation in the X- or Y-space.

- ***T-scores*** are the new coordinates of the data points in the X-space, computed in such a way that they capture the part of the structure in X which is most predictive for Y.
- ***U-scores*** summarize the part of the structure in Y which is explained by X along a given model component. (***Note:*** they do not exist in PCR!)

The relationship between t- and u-scores is a summary of the relationship between X and Y along a specific model component. For diagnostic purposes, this relationship can be visualized using the **X-Y Relation Outliers** plot.

PLS Loadings

The PLS loadings used in The Unscrambler express how each of the X- and Y-variables is related to the model component summarized by the t-scores. It follows that the loadings will be interpreted somewhat differently in the X- and Y-space.

- ***P-loadings*** express how much each X-variable contributes to a specific model component, and can be used exactly the same way as PCA loadings. Directions determined by the projections of the X-variables are used to interpret the meaning of the location of a projected data point on a t-score plot in terms of variations in X.
- ***Q-loadings*** express the direct relationship between the Y-variables and the t-scores. Thus, the directions determined by the projections of the Y-variables (by means of the q-loadings) can be used to interpret the meaning of the location of a projected data point on a t-score plot in terms of sample variation in Y.

The two kinds of loadings can be plotted on a single graph to facilitate the interpretation of the t-scores with regard to directions of variation both in X and Y. It must be pointed out that, contrary to PCA loadings, PLS loadings are not normalized, so that p- and q-loadings do not share a common scale. Thus, their directions are easier to interpret than their lengths, and the directions should only be interpreted provided that the corresponding X- or Y-variables are sufficiently taken into account (which can be checked using explained or residual variances).

PLS Loading Weights

Loading weights are specific to PLS (they have no equivalent in PCR) and express how the information in each X-variable relates to the variation in Y summarized by the u-scores. They are called loading ***weights*** because they also express, in the PLS algorithm, how the t-scores are to be computed from the X-matrix to obtain an orthogonal decomposition. The loading weights are normalized, so that their lengths can be interpreted as well as their directions. Variables with large loading weight values are important for the prediction of Y.

More Details About Regression Methods

It may be somewhat confusing to have a choice between three different methods that apparently solve the same problem: fit a model in order to approximate Y as a linear function of X.

The sections that follow will help you compare the three methods and select the one which is best adapted to your data and requirements.

MLR vs. PCR vs. PLS

MLR has the following properties and behavior:

- The number of X-variables must be smaller than the number of samples;
- In case of collinearity among X-variables, the b-coefficients are not reliable and the model may be unstable;
- MLR tends to overfit when noisy data is used.

PCR and PLS are projection methods, like PCA.

Model components are extracted in such a way that the first PC conveys the largest amount of information, followed by the second PC, etc. At a certain point, the variation modeled by any new PC is mostly noise. The optimal number of PCs - modeling useful information, but avoiding overfitting - is determined with the help of the residual variances.

PCR uses MLR in the regression step; a PCR model using all PCs gives the same solution as MLR (and so does a PLS1 model using all PCs).

If you run MLR, PCR and PLS1 on the same data, you can compare their performance by checking validation errors (Predicted vs. Measured Y-values for validation samples, RMSEP).

It can also be noted that both MLR and PCR only model one Y-variable at a time.

The ***difference between PCR and PLS*** lies in the algorithm. PLS uses the information lying in both X and Y to fit the model, switching between X and Y iteratively to find the relevant PCs. So PLS often needs fewer PCs to reach the optimal solution because the focus is on the prediction of the Y-variables (not on achieving the best projection of X as in PCA).

How To Select Regression Method

If there is more than one Y-variable, ***PLS2*** is usually the best method if you wish to interpret all variables simultaneously. It is often argued that ***PLS1*** or ***PCR*** give better prediction ability. This is usually true if there are strong non-linearities in the data, in which case modeling each Y-variable separately according to its own non-linear features might perform better than trying to build a common model for all Ys. On the other hand, if the Y-variables are somewhat noisy, but strongly correlated, PLS2 is the best way to model the whole information and leave noise aside.

The difference between PLS1 and PCR is usually quite small, but PLS1 will usually give results comparable to PCR-results using fewer components.

MLR should only be used if the number of X-variables is low and there are only small correlations among them.

Formal ***tests of significance*** for the regression coefficients are well-known and accepted for MLR. If you choose PCR or PLS, you may still check the stability of your results and the significance of the regression coefficients with Martens' Uncertainty Test.

How To Interpret Regression Results

Once a regression model is built, you have to ***diagnose*** it, i.e. assess its quality, before you can start ***interpreting*** the relationship between X and Y. Finally, your model will be ready for use for ***prediction*** once you have thoroughly checked and refined it.

The various types of results from MLR, PCR and PLS regression models are presented and their interpretation is roughly described in the above chapter Main Results Of Regression p.111.

You may find more about the interpretation of projection results (scores and loadings) and variance curves for PCR and PLS in the corresponding chapters covering PCA:

- Interpretation of variances p. 101
- Interpretation of scores and loadings p. 102

How To Detect Non-linearities (Lack Of Fit) In Regression

Different types of ***residual plots*** can be used to detect non-linearities or lack of fit. If the model is good, the residuals should be randomly distributed, and these plots should be free from systematic trends.

The most useful residual plots are the **Y-residuals vs. predicted Y** and **Y-residuals vs. scores** plots. Variable residuals can also sometimes be useful.

The PLS **X-Y Relation Outliers** plot is also a powerful tool to detect non-linearities, since it shows the shape of the relationship between X and Y along one specific model component.

How To Detect Outliers In Regression

As in PCA, outliers can be detected using score plots, residuals and leverages, but some of them in a slightly different way.

What is an Outlier?

Lookup Chapter "How To Detect Outliers in PCA" p. 101.

Outliers in Regression

In regression, there are many ways for a sample to be classified as an outlier. It may be outlying according to the ***X-variables only***, or to the ***Y-variables only***, or to ***both***. It may also not be an outlier for either separate set

of variables, but become an outlier when you consider the ***(X,Y) relationship***. In the latter case, the **X-Y Relation Outliers** plot (only available for PLS) is a very powerful tool showing the (X,Y) relationship and how well the data points fit into it.

Use of Residuals to Detect Outliers

You can use the residuals in several ways. For instance, first use ***residual variance pr sample***. Then use a ***variable residual*** plot for the samples showing up with large squared residual in the first plot. The first of the two plots is used for indicating samples with outlying variables, while the latter plot is used for a detailed study for each of these samples. In both cases, points located **far from the zero line** indicate outlying samples or variables.

Use of Leverages to Detect Outliers

The ***leverages*** are usually plotted versus sample number. Samples showing up with much larger leverage than the rest of the samples are outliers and may have had a strong influence on the model, which should be avoided.

For calibration samples, it is also natural to use an ***influence plot***. This is a plot of squared residuals (either X or Y) versus leverages. Samples with both large residuals and large leverage can then be detected. These are the samples with the strongest influence on the model, and can be harmful.

You can nicely combine those features with the double plot for ***influence*** and ***Y-residuals vs. predicted Y***.

Multivariate Regression in Practice

In practice, building and using a regression model consists of several steps:

1. Choose and implement an appropriate ***pre-processing*** method (see Chapter Re-formatting and Pre-processing p. **Feil! Bokmerke er ikke definert.**);
2. Build the model: ***calibration*** fits the model to the available data, while ***validation*** checks the model for new data;
3. Choose the ***number of components*** to interpret (for PCR and PLS), according to calibration and validation variances;
4. ***Diagnose*** the model, using outlier warnings, variance curves (for PCR and PLS), X-Y relation outliers (for PLS), Predicted vs. Measured;
5. ***Interpret*** the loadings and scores plots (for PCR and PLS), the loading weights plots (for PLS), Uncertainty Test results (for PCR and PLS – see Chapter Uncertainty Testing with Cross Validation p. 123), the B-coefficients, optionally the response surface
6. ***Predict*** response values for new data (optional).

The sections that follow list menu options and dialogs for data analysis and result interpretation using Regression. For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo's web site www.camo.com/TheUnscrambler/Appendices .

Run A Regression

When your data table is displayed in the ***Editor***, you may access the **Task** menu to run a suitable analysis – here, Regression.

Note: If the data table displayed in the ***Editor*** is a 3-D table, the **Task - Regression** menu option described hereafter allows you to perform three-way data modeling with nPLS. For more details concerning that application, lookup Chapter Three-way Data Analysis in Practice.

- **Task - Regression**: Run a Regression on the current data table

Save And Retrieve Regression Results

Once the regression model has been computed according to your specifications, you may either View the results right away, or Close (and Save) your regression result file to be opened later in the ***Viewer***.

Save Result File from the Viewer

- **File - Save**: Save result file for the first time, or with existing name
- **File - Save As**: Save result file under a new name

Open Result File into a new Viewer

- **File - Open**: Open any file or just lookup file information
- **Results - Regression**: Open regression result file or just lookup file information, warnings and variances
- **Results - All**: Open any result file or just lookup file information, warnings and variances

View Regression Results

Display regression results as plots from the ***Viewer***. Your regression results file should be opened in the Viewer; you may then access the **Plot** menu to select the various results you want to plot and interpret.

From the **View**, **Edit** and **Window** menus you may use more options to enhance your plots and ease result interpretation.

How To Plot Regression Results

- **Plot - Regression Overview**: Display the 4 main regression plots
- **Plot - Variances and RMSEP**: Plot variance curves (PCR, PLS)
- **Plot - Sample Outliers**: Display 4 plots for diagnosing outliers
- **Plot - X-Y Relation Outliers**: Display t vs. u scores along individual PCs (PLS)
- **Plot - Predicted vs Measured**: Display plot of predicted Y values against actual Y values
- **Plot - Scores and Loadings**: Display scores and loadings separately or as a bi-plot (PCR, PLS)
- **Plot - Scores**: Plot scores along selected PCs (PCR, PLS)
- **Plot - Loadings**: Plot loadings along selected PCs (PCR, PLS)
- **Plot - Loading Weights**: Plot loading weights along selected PCs (PLS)
- **Plot - Residuals**: Display various types of residual plots
- **Plot - Leverage**: Plot sample leverages
- **Plot - Important Variables**: Display 2 plots to detect most important variables (PCR, PLS)
- **Plot - Regression Coefficients**: Plot regression coefficients
- **Plot - Regression and Prediction**: Display Predicted vs. Measured and Regression coefficients
- **Plot - Response Surface**: Plot predicted Y values as a function of 2 or 3 X-variables
- **Plot - Analysis of Variance**: Display ANOVA table (MLR)

How To Display Uncertainty Results

- **View - Hotelling T2 Ellipse**: Display Hotelling T^2 ellipse on a score plot
- **View - Uncertainty Test - Stability Plot**: Display stability plot for scores or loadings
- **View - Uncertainty Test - Uncertainty Limits**: Display uncertainty limits on regression coefficients plot
- **View - Correlation Loadings**: Change a loading plot to display correlation loadings

For more options allowing you to re-format your plots, navigate along PCs, mark objects etc., look up chapter View PCA Results p. 103. All the menu options shown there also apply to regression results.

Run New Analyses From The Viewer

In the ***Viewer***, you may not only **Plot** your regression results; the **Edit - Mark** menu allows you to mark samples or variables that you want to keep track of (they will then appear marked on all plots), while the **Task - Recalculate…** options make it possible to re-specify your analysis without leaving the viewer.

Check that the currently active subview contains the right type of plot (samples or variables) before using **Edit - Mark**.

Application example

If you have used the Uncertainty Test option when computing your PCR or PLS model, you may ***mark all significant X-variables*** on a loading plot, then ***recalculate*** the model with only the ***marked*** X-variables.

The new model will usually ***fit as well*** as the original and ***validate better*** when variables with no significant contribution to the prediction of Y are removed.

How To Keep Track of Interesting Objects

- **Edit - Mark - One By One**: Mark samples or variables individually on current plot
- **Edit - Mark - With Rectangle**: Mark samples or variables by enclosing them in a rectangular frame (on current plot)
- **Edit - Mark - Significant X-variables Only**: Mark significant X-variables (only available if you used uncertainty testing)
- **Edit - Mark - Outliers Only**: Mark automatically detected outliers
- **Edit - Mark - Test Samples Only**: Mark test samples (only available if you used test set validation)
- **Edit - Mark - Evenly Distributed Samples Only**: Mark a subset of samples which evenly cover your data range

How To Remove Marking

- **Edit - Mark - Unmark All**: Remove marking for all objects of the type displayed on current plot

How To Reverse Marking

- **Edit - Mark - Reverse Marking**: Exchange marked and unmarked objects on the plot

How To Re-specify your Analysis

- **Task - Recalculate with Marked**: Recalculate model with only the marked samples / variables

- **Task - Recalculate without Marked**: Recalculate model without the marked samples / variables
- **Task - Recalculate with Passified Marked**: Recalculate model with marked variables weighted down using “Passify”
- **Task - Recalculate with Passified Unmarked**: Recalculate model with unmarked variables weighted down using “Passify”

Extract Data From The Viewer

From the ***Viewer***, use the **Edit - Mark** menu to mark samples or variables that you have reason to single out, e.g “significant X-variables” or “outlying samples”, etc.

A former chapter “Extract Data From The Viewer” p. 105 describes the options available for PCA Results. All the menu options shown there also apply to regression results.

Validate A Model

Check how well your PCA or regression model may apply to new data of the same kind as your model is based upon.

Principles of Model Validation

This chapter presents the purposes and principles of model validation in multivariate data analysis.

In order to make this presentation as general as possible, we will focus on the case of a ***regression*** model. However, the same principles apply to PCA.

If you are interested in the validation of PCA results:

- disregard any mention of "Y-variables";
- disregard the sections on RMSEP;
- and replace the word "predict" with "fit".

What Is Validation?

Validating a model means checking how well the model will perform on new data.

A regression model is usually made to do predictions in the future. The validation of the model estimates the uncertainty of such future predictions. If the uncertainty is reasonably low, the model can be considered valid.

The same argument applies to a descriptive multivariate analysis such as PCA: If you want to extrapolate the correlations observed in your data table to future, similar data, you should check whether they still apply for new data.

In The Unscrambler, three methods are available to estimate the prediction error: ***test set validation***, ***cross validation*** and ***leverage correction***.

Test Set Validation

Test set validation is based on testing the model on a subset of the available samples, which will not be present in the computations of the model components.

The global data table is split into two subsets:

1. The ***calibration set*** contains all samples used to compute the model components, using X- and Y-values;
2. The ***test set*** contains all the remaining samples, for which X-values are fed into the model once a new component has been computed. Their predicted Y-values are then compared to the observed Y-values, yielding a prediction residual that can be used to compute a validation residual variance or an RMSEP.

How To Select A Test Set

A test set should contain 20-40% of the full data table. The calibration and test set should in principle cover the same population of samples as well as possible. Samples which can be considered to be replicate measurements should not be present in both the calibration and test set.

There are several ways to select test sets:

- ***Manual selection*** is recommended since it gives you full control over the selection of a test set;
- ***Random selection*** is the simplest way to select a test set, but leaves the selection to the computer;
- ***Group selection*** makes it possible for you to specify a set of samples as test set by selecting a value or values for one of the variables. This should only be used under special circumstances. An example of such a situation is a case where there are two true replicates for each data point, and a separate variable indicates which replicate a sample belongs to. In such a case, one can construct two groups according to this variable and use one of the sets as test set.

Cross Validation

With ***cross validation***, the same samples are used both for model estimation and testing. A few samples are left out from the calibration data set and the model is calibrated on the remaining data points. Then the values for the left-out samples are predicted and the prediction residuals are computed. The process is repeated with another subset of the calibration set, and so on until every object has been left out once; then all prediction residuals are combined to compute the validation residual variance and RMSEP.

Several versions of the cross validation approach can be used:

- ***Full cross validation*** leaves out only one sample at a time; it is the original version of the method;
- ***Segmented cross validation*** leaves out a whole group of samples at a time;
- ***Test-set switch*** divides the global data set into two subsets, each of which will be used alternatively as calibration set and as test set.

Leverage Correction

Leverage correction is an approximation to cross validation that enables prediction residuals to be estimated without actually performing any prediction. It is based on an equation that is valid for MLR, but is only an approximation for PLS and PCR.

According to this equation, the prediction residual equals
(calibration residual) divided by (1 - sample leverage).

All samples with low leverage (i.e. low influence on the model) will have estimated prediction residuals very close to their calibration residuals (the leverage being close to zero). For samples with high leverage, the calibration residual will be divided by a smaller number, thus giving a much larger estimated prediction residual.

Validation Results

The simplest and most efficient measure of the uncertainty on future predictions is the ***RMSEP*** (Root Mean Square Error of Prediction). This value (one for each response) tells you the average uncertainty that can be expected when predicting Y-values for new samples, expressed in the same units as the Y-variable. The results of future predictions can then be presented as "predicted values ± 2•RMSEP". This measure is valid provided that the new samples are similar to the ones used for calibration, otherwise, the prediction error might be much higher.

Validation residual and explained variances are also computed in exactly the same way as calibration variances, except that prediction residuals are used instead of calibration residuals. Validation variances are used, as in PCA, to find the optimum number of model components. When validation residual variance is minimal, RMSEP also is, and the model with an optimal number of components will have the lowest expected prediction error.

RMSEP can be compared with the precision of the reference method. Usually you cannot expect RMSEP to be lower than twice the precision.

When To Use Which Validation Method

Properties of Test Set Validation

Test set validation can be used if there are many samples in the data table, for instance more than 50.

It is the most "objective" validation method, since the test samples do not influence the calibration of the model.

Properties of Cross Validation

Cross validation represents a more efficient way of utilizing the samples if the number of samples is small or moderate, but is considerably slower than test set validation.

Segmented cross validation is faster, but usually, ***full cross validation*** improves the relevance and power of the analysis. If you use segmented cross validation, make sure that all segments contain unique information, i.e. samples which can be considered as replicates of each other should not be present in different segments.

The major advantage of cross validation is that it allows for the ***jack-knifing*** approach on which Martens' Uncertainty Test is based. This provides you with significance testing for PCR and PLS results. For more information, see Uncertainty Testing With Cross Validation hereafter.

Properties of Leverage Correction

Leverage correction for projection methods should only be used in an ***early stage of the analysis*** if it is very important to obtain a quick answer. In general it gives more "optimistic" results than the other validation methods and can sometimes be highly overoptimistic.

Sometimes, especially for small data tables, leverage correction can give apparently reasonable results, while cross validation fails completely. In such cases, the "reasonable" behavior of the leverage correction can be an artifact and cannot be trusted. The reason why such cases are difficult is that there is too little information for estimation of a model and each sample is "unique". Therefore all known validation methods are doomed to fail.

For MLR, leverage correction is strictly equivalent to (and much faster than) full cross validation.

Uncertainty Testing With Cross Validation

Users of multivariate modeling methods are often uncertain when interpreting models. Frequently asked questions are:

- Which variables are ***significant***?

- Is the model ***stable***?

- Why is there a ***problem***?

Dr Harald Martens has developed a new and unique method for uncertainty testing, which gives ***safer interpretation of models***. The concept for uncertainty testing is based on cross validation, Jack-knifing and stability plots. This chapter introduces how Martens' Uncertainty Test works and shows how you use it in The Unscrambler through an application.

The following sections will present the method with a non-mathematical approach.

How Does Martens' Uncertainty Test Work?

The test works with PLS, PCR or PCA models with cross validation, choosing full cross validation or segmented cross validation as is appropriate for the data. When you have chosen the optimal number of PLS- or Principal Components (PCs), tick **Uncertainty Test** in The Unscrambler modeling dialog box.

Under cross validation, a number of ***sub-models*** are created. These sub-models are based on all the samples that were not kept out in the cross validation segment. For every sub-model, a set of model parameters: B-coefficients, scores, loadings and loading weights are calculated. Variations over these sub-models will be estimated so as to assess the stability of the results.

In addition a total model is generated, based on ***all*** the samples. This is the model that you will interpret.

Uncertainty of Regression Coefficients

For each variable we can calculate the difference between the B-coefficient B_i in a sub-model and the B_{tot} for the total model. The Unscrambler takes the sum of the squares of the differences in all sub-models to get an expression of the variance of the B_i estimate for a variable.

With a t-test the ***significance*** of the estimate of B_i is calculated. Thus the resulting regression coefficients can be presented with ***uncertainty limits*** that correspond to 2 Standard Deviations under ideal conditions. Variables with uncertainty limits that ***do not cross the zero line*** are significant variables.

Uncertainty of Loadings and Loading Weights

The same can be done for the other model parameters, but there is a rotational ambiguity in the latent variables of bilinear models. To be able to compare all the sub-models correctly, Dr. Martens has chosen to rotate them. Therefore we can also get uncertainty limits for these parameters.

Stability Plots

The results of all these calculations can also be visualized as ***stability plots*** in scores, loadings, and loading weights plots. Stability plots can be used to understand the influence of specific samples and variables on the model, and explain for example why a variable with a large regression coefficient is not significant. This will be illustrated in the example that follows (see Application Example).

Easier to Interpret Important Variables in Models with Many Components

Models with ***many components***, three, four or more, may be ***difficult to interpret***, especially if the first PCs do not explain much of the variance.

For instance, if each of the first 4-5 PCs explain 15-20%, the PC1/PC2 plot is not enough to understand which are the most important variables.

In such cases, Martens' automatic uncertainty test shows you the significant variables in the many-component model and interpretation is far easier.

Remove Non-Significant Variables for more Robust Models

Variables that are non-significant display non-structured variation, i.e. noise. When you remove them, the resulting model will be more ***stable*** and ***robust*** (i.e. less sensitive to noise). Usually the prediction error decreases too.

Therefore, after identifying the significant variables by using the automatic marking based on Martens' test, use The Unscrambler function **Recalculate with Marked** to make a new model and check the improvements.

Application Areas

1. ***Spectroscopic calibrations*** work better if you remove noisy wavelengths.
2. Some models may be improved by adding ***interations and squares*** of the variables, and The Unscrambler has a feature to do this automatically. However, many of these terms are irrelevant. Apply Martens' uncertainty test to identify and keep only the significant ones.

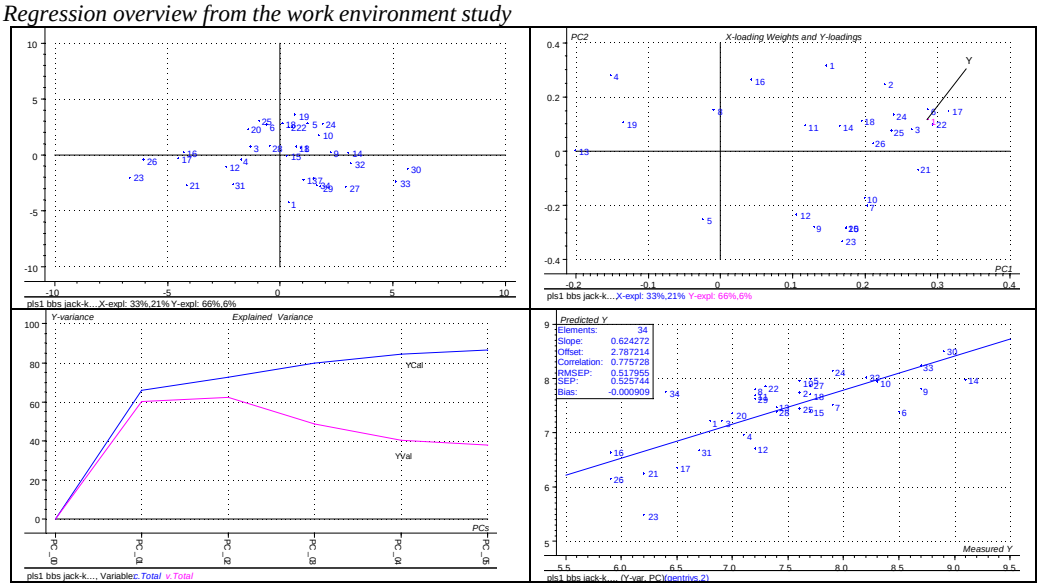
Application Example

In a ***work environment study***, we used PLS1 to model 34 data samples corresponding to 34 departments in a company. The data was collected from a questionnaire about feeling good at work (Y), modeled from 26 questions (X1, X2, … X26) about repetitive tasks, inspiration from the boss, helpful colleagues, positive feedback from the boss, etc.

The model has 2 PCs assessed by ***full cross validation*** and Uncertainty Test. Thus the cross validation has created 34 sub-models, where 1 sample has been left out in each.

The Unscrambler regression overview shown in the figure below contains a Score plot (PC1-PC2), the X-Loading Weights and Y-loadings plot (PC1-PC2), the explained variance and the Predicted vs. Measured plot for 2 PCs for this PLS1 regression model.

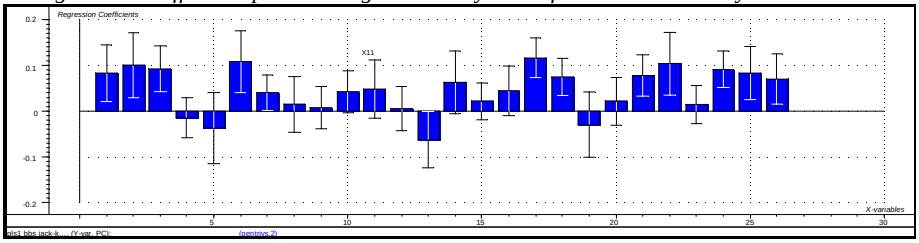
Regression overview from the work environment study



Work Environment Study: Significant Variables

When plotting the regression coefficients we can also plot the uncertainty limits as shown below.

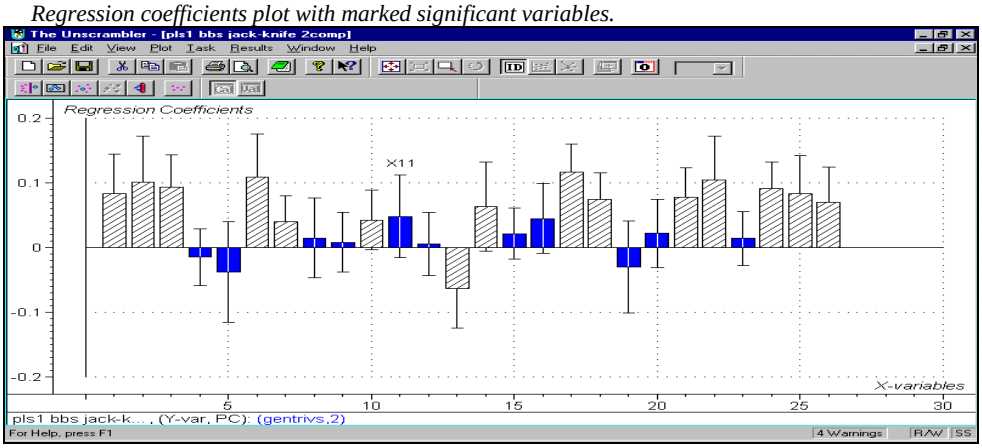
Regression coefficients plot showing uncertainty limits from the Uncertainty Test.



Variable X11's regression coefficient has uncertainty limits crossing the zero line: it is not significant.

The automatic function "*Mark significant variables*" shows clearly which variables have a significant effect on Y (see figure below).

Regression coefficients plot with marked significant variables.

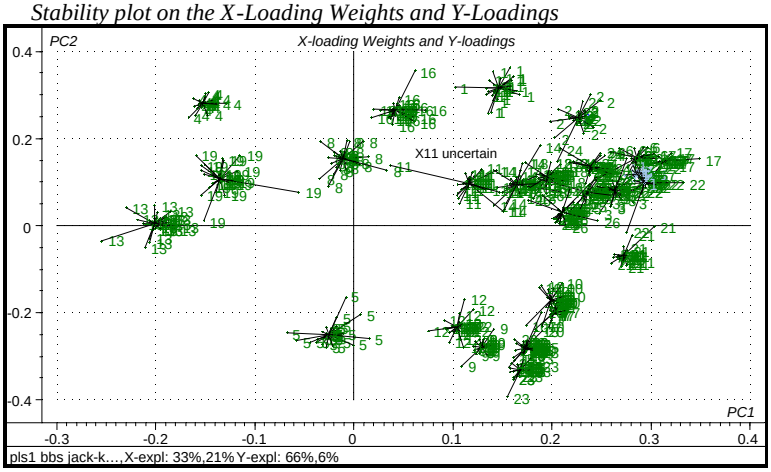


15 X-variables out of 26 are significant. X11 ("Do you get help from your colleagues?") is not significant, even though its B-coefficient is not among the smallest. How come?

Work Environment Study: Stability in Loading Weights Plots

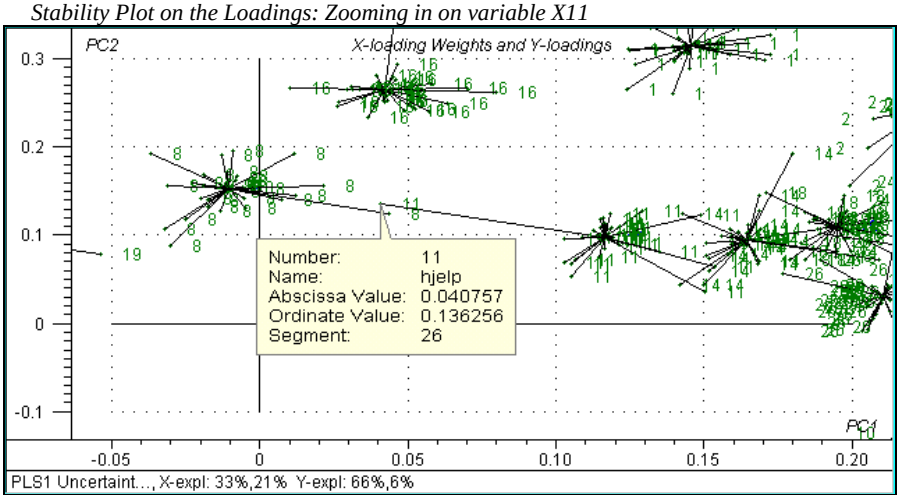
By clicking the icon for Stability plot when studying Loading Weights, we get the picture shown below:

Stability plot on the X-Loading Weights and Y-Loadings



For each variable you see a swarm of its loading weights in each sub-model. There are 26 such X-loading weights swarms. In the ***middle*** of each swarm you see the loading weight for the variable in the total model. They should lie close together. Usually the uncertainty is larger (the spread is larger in the swarm) for variables close to the origin, i.e. these variables are non-significant.

Stability Plot on the Loadings: Zooming in on variable X11

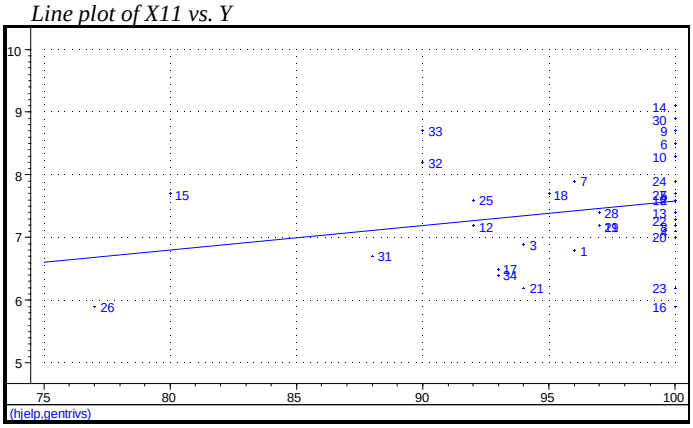


If a variable has a sub-loading far away from the rest in its swarm, then this variable is strongly influenced by one of the sub-models. The segment information on the figure above indicates that sub-model 26 (or "segment" 26 as shown in the pop-up information) has a large influence on variable X11.

Individual samples can be very ***influential*** when included in a model. In segment 26, where sample 26 was kept out, the sub-loading weight for variable X11 is very different from the sub-loading weights obtained from all other sub-models, where sample 26 was included. Probably this sample has an extreme value for variable X11, so the distribution is skewed. Therefore the estimate of the loading weight for variable X11 is uncertain, and it becomes non-significant.

We can verify the extreme value of sample 26 by plotting X11 versus Y as shown below:

Line plot of X11 vs. Y

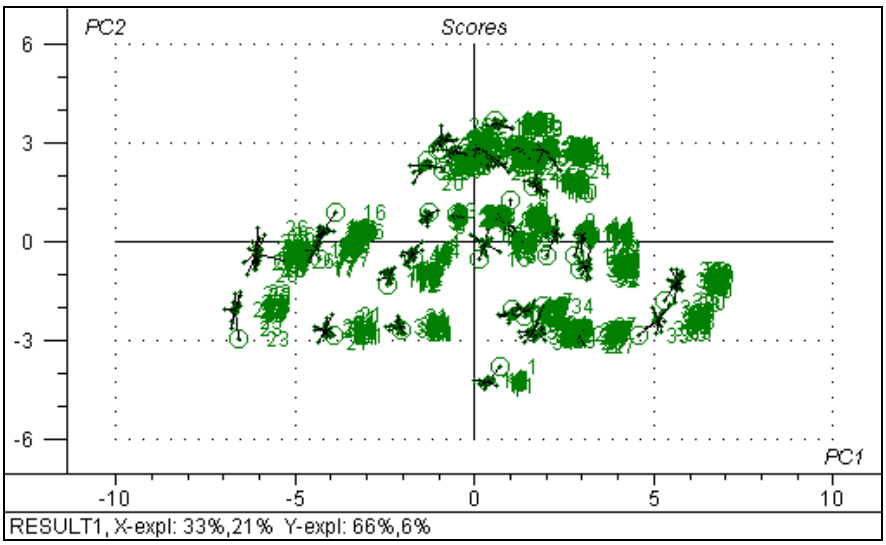


Only two departments (15 and 26) consider their colleagues not being helpful, so these two samples influence the sub-models strongly and twist them. Without these two samples, variable X11 would have a very small variation and the model would be different. Sample 26 clearly drags the regression line down. By removing it you would get a fairly horizontal line, i.e. no relationship at all between X11 and Y.

Work Environment Study: Stability in Scores Plots

The figure below shows the plot obtained by clicking the icon for Stability plot when studying scores.

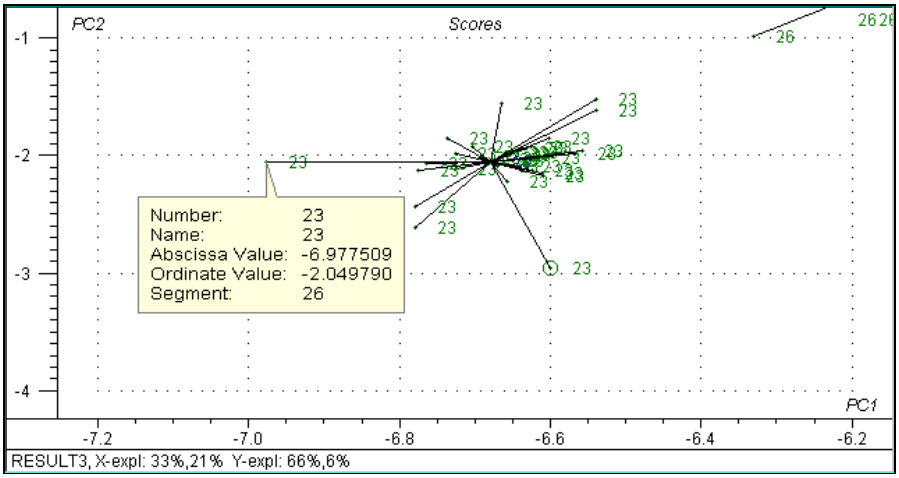
Stability Plot on the Scores



For each sample you see a swarm of its scores from each sub-model. There are 34 sample swarms. In the ***middle*** of each swarm you see the score for the sample in the total model. The ***circle*** shows the projected or rotated score of the sample in the sub-model where it was left out.

The next figure presents a ***zooming on sample 23***. The sub-score marked with a circle corresponds to the sub-model where sample 23 was kept out. The segment information displayed on the figure points towards the sub-score for sample 23 when sample 26 was kept out. Here again, we observe the influence of sample 26 on the model.

Stability Plot on the scores: Zooming in on sample 23



If a given sample is far away from the rest of the swarm, it means that the sub-model without this sample is very different from the other sub-models. In other words, this sample has influenced all other sub-models due to its uniqueness.

In the work environment example, from looking at the global picture from the stability score plot we can conclude that all samples seem OK and the model seems robust.

More Details About The Uncertainty Test

One of the critiques towards PLS regression has been the lack of significance of the model parameters. Many years of experience have given "rules of thumb" of how to find which variables are significant. However, these "rules of thumb" do not apply in all cases, and the users still see the need for easy interpretation and guidance in these matters. The data analysis must give reasonable protection against wishful thinking based on spurious effects in the data. To be effective, such statistical validation must be easily understood by its user.

The modified Jack-knifing method implemented in The Unscrambler has been invented by Harald Martens, and was published in Food Quality and Preference (1999). Its details are presented hereafter.

Note: To understand this chapter, you need basic knowledge about the purposes and principles of chemometrics. If you have never worked with multivariate data analysis before, we strongly recommend that you read about it in the chapters about PCA and regression before proceeding with this chapter.

See the Application Example above for details of how to use the Uncertainty Test results in practice.

New Assessment of Model Parameters

The cross validation assessment of the predictive validity is here extended to uncertainty assessment of the individual model parameters: In each cross validation segment **m**=1,2,...,M a perturbed version of the structure model described is obtained.

We refer to the Method References chapter, which is available as a .PDF file from CAMO's web site www.camo.com/TheUnscrambler/Appendices , for the mathematical details of PCA, PCR and PLS regression.

Each ***perturbed model*** is based on all the objects except one or more objects which were kept 'secret' in this cross validation segment **m**.

If a perturbed segment model ***differs*** greatly from the common model, based on all the objects, it means that the object(s) kept 'secret' in this cross validation segment have significantly affected the common model. These left out objects caused some ***unique pattern of variation*** in the model parameters. Thus, a plot of how the model parameters are perturbed when different objects are kept 'secret' in the different cross validation segments **m**=1,2,...,M shows the ***robustness*** of the common model against peculiarities in the data of individual objects or segments of objects.

These perturbations may be inspected graphically in order to acquire a general impression of the ***stability*** of the parameter estimates, and to identify dominating ***sources of model instability***. Furthermore, they may also be summarized to yield estimates of the variance/covariance of the model parameters.

This is often called ***"jack-knifing"***. It will here be used for two purposes:

3. Elimination of ***useless variables***, based on the linear parameters **B**;
4. ***Stability assessment*** of the bilinear structure parameters **T** and [**P'**, **Q'**].

Rotation of Perturbed Models

It is also important to be able to assess the bilinear score and loading parameters. However, the bilinear structure model has a related rotational ***ambiguity*** in the latent variables that needs to be corrected for in the jack-knifing. Only then is it meaningful to assess the perturbations of scores $\mathbf{T}_m$ and loadings $\mathbf{P}_m$ and $\mathbf{Q}_m$ in cross validation model segment # *m*. Any invertible matrix $\mathbf{C}_m$ (*AxA*) satisfies the relationships:

$$\mathbf{T}_m\left[\mathbf{P}_m', \mathbf{Q}_m'\right] = \mathbf{T}_m\mathbf{C}_m\mathbf{C}_m^{-1}\left[\mathbf{P}_m', \mathbf{Q}_m'\right]$$

Therefore, the individual models *m=1,2,...,M* may be rotated, e.g. towards a common model:

$$\mathbf{T}_{(m)} = \mathbf{T}_m \hat{\mathbf{C}}_m$$
$$[\mathbf{P}', \mathbf{Q}']_{(m)} = \hat{\mathbf{C}}_m^{-1} [\mathbf{P}_m', \mathbf{Q}_m']$$

After rotation, the rotated parameters $\mathbf{T}_{(m)}$ and $[\mathbf{P}', \mathbf{Q}']_{(m)}$ may be compared to the corresponding parameters from the common model **T** and [**P', Q'**]. The perturbations may then be written as $(\mathbf{T}_{(m)} - \mathbf{T})g$ and or $([\mathbf{P}', \mathbf{Q}']_{(m)} - [\mathbf{P}', \mathbf{Q}'])g$ for the scores and the loadings, respectively, where *g* is a scaling factor (here: *g=1*).

In the implemented code, an ***orthogonal Procrustes rotation*** is used. The same rotation principle is also applied for the loading weights, **W**, where a separate rotation matrix is computed for **W**. The uncertainty estimates for **P, Q** and **W** are estimated in the same manner as for **B** below.

Eliminating Useless Variables

On the basis of such jack-knife estimates of the uncertainty of the model parameters, useless or unreliable X-or Y-variables may be eliminated automatically, in order to simplify the final model and making it more reliable. The following part describes the cross validation / jack-knifing procedure:

When cross validation is applied in regression, the optimal rank A is determined based on prediction of kept-out objects (samples) from the individual models. The approximate uncertainty variance of the PCR and PLS regression coefficients B can be estimated by jack-knifing

$$S^2B = \sum_{m=1}^{M} ((B - B_m)g)^2$$

where

S^2B (K x J) = estimated uncertainty variance of B

B (K x J) = the regression coefficient at the cross validated rank A using all the N objects,

B_m (K x J) = the regression coefficient at the rank A using all objects except the object(s) left out in cross validation segment m

g = scaling coefficient (here: g=1).

Significance Testing

When the variances for **B**, **P**, **Q**, and **W** have been estimated, they can be utilized to find significant parameters.

As a rough significance test, a Student's ***t-test*** is performed for each element in **B** relative to the square root of its estimated uncertainty variance $\mathbf{S^2B}$, giving the significance level for each parameter. In addition to the significance for **B**, which gives the overall significance for a specific number of components, the significance levels for **Q** are useful to find in which components the Y-Variables are modeled with statistical relevance.

Model Validation in Practice

The sections that follow list menu options, dialogs and plots for model validation. For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo's web site www.camo.com/TheUnscrambler/Appendices .

How To Validate A Model

In The Unscrambler, validation is always automatically included in model computation. However, what matters most is the choice of a relevant validation method for your case, and the configuration of its parameters.

The ***general validation procedure*** for PCA and Regression is as follows:

1. Build a ***first model*** with leverage correction or segmented cross validation – the computations will go faster. Allow for a large number of PCs. Cross validation is recommended if you wish to apply Martens' Uncertainty Test.
2. ***Diagnose*** the first model with respect to outliers, non linearities, any other abnormal behavior. Take advantage of the variety of diagnostic tools available in The Unscrambler: variance curves, automatic warnings, scores and loadings, stability plots, influence plot, X-Y relation outliers plot, etc.
3. ***Investigate*** and ***fix*** problems (correct errors, apply transformations etc.)
4. Check ***improvements*** by building new model.
5. For ***regression*** only: validate intermediate model with a full cross validation, using Uncertainty Testing, then do ***variable selection*** based on significant regression coefficients.
6. ***Validate final model*** with a proper method (test set or full cross validation).
7. ***Interpret*** final model (sample properties, variable relationships, etc.). Check RMSEP for regression models.

Analysis and Validation Procedures

- **Task - PCA**: Starts the PCA dialog where you may choose a validation method and further specify validation details
- **Task - Regression**: Starts the Regression (PLS, PCR or MLR) dialog where you may choose a validation method and further specify validation details

Validation Dialogs

The following dialogs are accessed from the PCA dialog and Regression dialog at the **Task** stage:

- Cross Validation Setup
- Uncertainty Test
- Test Set Validation Setup

How To Display Validation Results

First, you should display your PCA or regression results as plots from the ***Viewer***. When your results file has been opened in the Viewer you may access the **Plot** and the **View** menus to select the various results you want to plot and interpret.

Open Result File into a new Viewer

- **Results - PCA**: Open PCA result file or just lookup file information, warnings and variances
- **Results - Regression**: Open regression result file or just lookup file information, warnings and variances
- **Results - All**: Open any result file or just lookup file information, warnings and variances

How To Display Validation Plots and Statistics

- **Plot - Variances and RMSEP**: Plot variance curves and estimated Prediction Error (PCA, PCR, PLS)
- **Plot - Predicted vs Measured**: Display plot of predicted Y values against actual Y values
- **View - Plot Statistics**: Display statistics (including RMSEP) on Predicted vs Measured plot
- **Plot - Residuals**: Display various types of residual plots
- **View - Source - Validation**: Toggle Validation results on/off on current plot
- **View - Source - Calibration**: Toggle Calibration results on/off on current plot
- **Window - Warning List**: Display general warnings issued during the analysis – among others related to validation

How To Display Uncertainty Test Results

First, you should display your PCA or regression results as plots from the ***Viewer***. When your results file has been opened in the Viewer you may access the **Plot** and the **View** menus to select the various results you want to plot and interpret.

How To Display Uncertainty Results

- **View - Hotelling T2 Ellipse**: Display Hotelling T^2 ellipse on a score plot
- **View - Uncertainty Test - Stability Plot**: Display stability plot for scores or loadings
- **View - Uncertainty Test - Uncertainty Limits**: Display uncertainty limits on regression coefficients plot
- **View - Correlation Loadings**: Change a loading plot to display correlation loadings

Make Predictions

Use an existing regression model to predict response values for new samples.

Principles of Prediction on New Samples

Prediction (computation of unknown response values using a regression model) is the purpose of most regression applications.

When Can You Use Prediction?

Prerequisites for prediction of response values on new samples for which X-values are available are the following:

- You need a ***regression model*** (MLR or PCR or PLS) which expresses the response variable or variables (Y) as a function of the X-variables;
- The model should have been ***calibrated*** on samples covering the region your new samples belong to, i.e. on ***similar samples*** (similarity being determined by the X-values);
- The model should also have been ***validated*** on samples covering the region your new samples belong to.

Note that model validation can only be considered successful if you have

- used a proper validation method (test set or cross validation);
- dealt with outliers in a proper way (not just removed all the samples which did not fit well);
- and obtained a value of RMSEP that you can live with.

How Does Prediction Work?

Prediction consists in feeding observed X-values for new samples into a regression model so as to obtain computed (predicted) Y-values.

As the next sections will show, this operation may be done in more than one way, at least for projection methods.

Prediction from an MLR Model

When you choose MLR as a regression method, there is only one way to compute predictions. It is based on the model equation, using the observed values for the ***X-variables***, and the ***regression coefficients*** (b_0, b_1, …, b_k) for the MLR model:

$$Y_{pred} = b_0 + b_1X_1 + ... + b_kX_k$$

This prediction method is simple and easy to understand. However it has a disadvantage, as we will see when we compare it to another approach presented in the next section.

Prediction from a PCR or PLS Model

If you choose PCR or PLS as a regression method, you may still compute predicted Y-values using X and the b-coefficients.

However, you can also take advantage of ***projection*** onto the model components to express predicted Y-values in a different way.

The PCR model equation can be written:

$$X = T \cdot P^T + E \text{ and } y = T \cdot b + f$$

and the PLS model equation:

$$X = T \cdot P^T + E \text{ and } Y = T \cdot B + F$$

In both these equations, we can see that Y is expressed as an ***indirect function*** of the X-variables, using the ***scores*** T.

The advantage of using the projection equation for prediction, is that when projecting a new sample onto the X-part of the model (this operation gives you the t-scores for the new sample), you simultaneously get a leverage value and an X-residual for the new sample that allow for ***outlier detection***.

A prediction sample with a high leverage and/or a large X-residual is a ***prediction outlier***. It cannot be considered as belonging to the same "population" as the samples your regression model is based on, and therefore you should not apply your model to the prediction of Y-values for such a sample.

Note: Using leverages and X-residuals, prediction outliers can be detected ***without any knowledge*** of the ***true value of Y***.

Prediction in The Unscrambler

Since projection allows for outlier detection, predictions done with a projection model (PCR, PLS) are safer than MLR predictions.

This is why The Unscrambler allows prediction only from PCR or PLS models, and provides you with tools to detect prediction outliers (which do not exist for MLR).

Main Results Of Prediction

The main results of prediction include **Predicted Y-values** and **Deviations**. They can be displayed as plots.

In addition, **warnings** are computed and help you detect outlying samples or individual values of some variables.

Predicted with Deviation

This plot shows the predicted Y-values for all samples, together with a ***deviation*** which expresses how similar the prediction sample is to the calibration samples used when building the model. The more similar, the smaller the deviation. Predicted Y-values for samples with high deviations cannot be trusted.

For each sample, the deviation (which is a kind of 95% confidence interval around the predicted Y-value) is computed as a function of the sample's ***leverage*** and its ***X-residual variance***. For more details, lookup Chapter "Deviation in Prediction" in the Method References chapter, which is available as a .PDF file from CAMO's web site www.camo.com/TheUnscrambler/Appendices .

Predicted vs. Reference

(Only available if reference response values are available for the prediction samples).

This is a 2-D scatter plot of Predicted Y-values vs. Reference Y-values. It has the same features as a Predicted vs. Measured plot.

Prediction in Practice

The sections that follow list menu options, dialogs and plots for prediction. For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo's web site www.camo.com/TheUnscrambler/Appendices .

Run A Prediction

In practice, prediction requires three operations:

1. Build and validate a ***regression model***, using PCR or PLS (see Chapter Multivariate Regression in Practice p. 116) – or, for three-way data, nPLS; ***save*** the final version of your model.
2. Collect ***X-values*** for ***new*** samples (for three-way data, you need both ***Primary*** and ***Secondary X-values***);
3. Run a ***prediction***, using the chosen regression model.

When your data table is displayed in the ***Editor***, you may access the **Task** menu to run a Prediction.

- **Task - Predict**: Run a prediction on some samples contained in the current data table

Save And Retrieve Prediction Results

Once the predictions have been computed according to your specifications, you may either View the results right away, or Close (and Save) your prediction result file to be opened later in the ***Viewer***.

Save Result File from the Viewer

- **File - Save**: Save result file for the first time, or with existing name
- **File - Save As**: Save result file under a new name

Open Result File into a new Viewer

- **File - Open**: Open any file or just lookup file information
- **Results - Prediction**: Open prediction result file or just lookup file information and warnings
- **Results - All**: Open any result file or just lookup file information, warnings and variances

View Prediction Results

Display prediction results as plots from the ***Viewer***. Your prediction results file should be opened in the Viewer; you may then access the **Plot** menu to select the various results you want to plot and interpret.

From the **View**, **Edit** and **Window** menus you may use more options to enhance your plots and ease result interpretation.

How To Plot Prediction Results

- **Plot - Prediction**: Display the prediction plots of your choice

PC Navigation Tool

Navigate up or down the PCs in your model along the vertical and horizontal axes of your plots:

- **View - Source - Previous Vertical PC**
- **View - Source - Next Vertical PC**
- **View - Source - Back to Suggested PC**
- **View - Source - Previous Horizontal PC**
- **View - Source - Next Horizontal PC**

More Plotting Options

- **Edit - Options**: Format your plot
- **Edit - Insert Draw Item**: Draw a line or add text to your plot
- **View - Plot Statistics**: Display plot statistics, including RMSEP, on your Predicted vs. Reference plot
- **View - Outlier List**: Display list of outlier warnings issued during the analysis for each PC, sample and/or variable
- **Window - Warning List**: Display general warnings issued during the analysis

How To Keep Track of Interesting Objects

- **Edit - Mark**: Several options for marking samples or variables

How To Re-specify your Prediction

- **Task - Recalculate with Marked**: Recalculate predictions with only the marked samples
- **Task - Recalculate without Marked**: Recalculate predictions without the marked samples

How To Display Raw Data

- **View - Raw Data**: Display the source data for the predictions in a slave Editor

How To Extract Raw Data (into New Table)

- **Task - Extract Data from Marked**: Extract data for only the marked samples
- **Task - Extract Data from Unmarked**: Extract data for only the unmarked samples

Classification

Use existing PCA models to build a SIMCA classification model, then classify new samples.

Principles of Sample Classification

This chapter presents the ***purposes*** of sample classification, and focuses on the major classification method available in The Unscrambler, which is ***SIMCA classification***.

There are alternative classification methods, like discriminant analysis which is widely used in the case of only two classes. A variant called PLS Discriminant Analysis will be briefly mentioned in the last section PLS Discriminant Analysis.

Purposes Of Classification

The main goal of classification is to reliably ***assign new samples to existing classes*** (in a given population).

Note that ***classification*** is not the same as ***clustering***.

You can also use classification results as a diagnostic tool:

- to distinguish among the most important variables to keep in a model (variables that "characterize" the population);
- or to find outliers (samples that are not typical of the population).

It follows that, contrary to regression, which predicts the values of one or several quantitative variables, ***classification*** is useful when the response is a category variable that can be interpreted in terms of several classes to which a sample may belong.

Examples of such situations are:

- Predicting whether a product meets ***quality requirements***, where the result is simply "Yes" or "No" (i.e. binary response).

- Modeling various close ***species of plants or animals*** according to their easily observable characteristics, so as to be able to decide whether new individuals belong to one of the modeled species.

- Modeling various ***diseases*** according to a set of easily observable symptoms, clinical signs or biological parameters, so as to help future diagnostic of those diseases.

SIMCA Classification

The classification method implemented in The Unscrambler is ***SIMCA*** (Soft Independent Modeling of Class Analogy).

SIMCA is based on making a PCA model for each class in the training set. Unknown samples are then compared to the class models and assigned to classes according to their analogy to the training samples.

Steps in Classification

Solving a classification problem requires two steps:

1. ***Modeling***: Build one separate model for each class;

2. ***Classifying*** new samples: Fit each sample to each model and decide whether the sample belongs to the corresponding class.

The modeling stage implies that you have identified enough samples as members of each class to be able to build a reliable model. It also requires enough variables to describe the samples accurately.

The actual classification stage uses significance tests, where the decisions are based on statistical tests performed on the object-to-model distances.

Making a SIMCA Model

SIMCA modeling consists in building ***one PCA model for each class***, which describes the structure of that class as well as possible. The optimal number of PCs should be chosen for each model separately, according to a suitable validation. Each model should be checked for possible outliers and improved if possible (like you would do for any PCA model).

Before using the models to predict class membership for new samples, you should also evaluate their specificity, i.e. whether the classes overlap or are sufficiently distant to each other. Specific tools, such as SIMCA results, are available for that purpose.

Classifying New Samples

Once each class has been modeled, and provided that the classes do not overlap too much, ***new samples*** can be fitted to (***projected*** onto) each model. This means that for each sample, new values for all variables are computed using the scores and loadings of the model, and compared to the actual values.

The residuals are then combined into a measure of the ***object-to-model distance***.

The scores are also used to build up a measure of the distance of the sample to the model center, called ***leverage***.

Finally, both object-to-model distance and leverage are taken into account to decide which class(es) the sample belongs to.

The classification decision rule is based on a classical statistical approach. If a sample belongs to a class, it should have a small distance to the class model (the ideal situation being “distance=0”). Given a new sample, you just need to compare its distance to the model to a class membership limit reflecting the probability distribution of object-to-model distances around zero.

Main Results of Classification

A SIMCA analysis gives you specific results in addition to the usual PCA results like scores, loadings, residuals.

These results are briefly listed hereafter, then detailed in the following sections.

Model Results

For each pair of models, **Model distance** between the two models is computed.

Variable Results

- **Modeling power** (of one variable in one model)
- **Discrimination power** (of one variable between two models).

Sample Results

- **Si** = object-to-model distance (of one sample to one model)
- **Hi** = leverage (of one sample to one model).

Combined Plots

- Si vs. Hi
- Cooman's plot.

Model Distance

This measure (which should actually be called "***model-to-model distance***") shows how different two models are from each other. It is computed from the results of fitting all samples from each class to their own model and to the other one.

The value of this measure should be compared to is 1 (distance of a model to itself). A model distance much larger than 1 (for instance, 3 or more) shows that the two models are quite different, which in turn implies that the two classes are likely to be well distinguished from each other.

Modeling Power

Modeling power is a measure of the influence of a variable over a given model. It is computed as
(1 - square root of (variable residual variance / variable total variance)).

This measure has values between 0 and 1; the closer to 1, the better that variable is taken into account in the class model, the higher the influence of that variable, and the more relevant it is to that particular class.

Discrimination Power

The ***discrimination power*** of a variable indicates the ability of that variable to discriminate between two models. Thus, a variable with a high discrimination power (with regard to two particular models) is very important for the differentiation between the two corresponding classes.

Like model distance, this measure should be compared to 1 (no discrimination power at all), and variables with a discrimination power higher than 3 can be considered quite important.

Sample-to-Model Distance (Si)

The ***sample-to-model distance*** is a measure of how far the sample lies from the modeled class. It is computed as the square root of the sample residual variance.

It can be compared to the overall variation of the class (called S0), and this is the basis of the statistical criterion used to decide whether a new sample can be classified as a member of the class or not. A small distance means that the sample is well described by the class model; it is then a likely class member.

Sample Leverage (Hi)

The ***sample leverage*** is a measure of how far the projection of a sample onto the model is from the class center, i.e. it expresses how different the sample is from the other class members, regardless of how well it can be described by the class model.

The leverage can take values between 0 and 1; the value is compared to a fixed limit which depends on the number of components and of calibration samples in the model.

Si vs. Hi

This plot is a graphical tool used to get a view of the sample-to-model distance (Si) and sample leverage (Hi) for a given model at the same time. It includes the class membership limits for both measures, so that samples can easily be classified according to that model by checking whether they fall inside both limits.

Cooman's Plot

This is an "***Si vs. Si***" plot, where the sample-to-model distances are plotted against each other for two models. It includes class membership limits for both models, so that you can see whether a sample is likely to belong to one class, or both, or none.

Outcomes Of A Classification

There are three possible outcomes of a classification:

1. Unknown sample belongs to **one** class;
2. Unknown sample belongs to **several** classes;
3. Unknown sample belongs to **none** of the classes.

The first case is the easiest to interpret.

If the classes have been modeled with enough precision, the second case should not occur (no overlap). If it does occur, this means that the class models might need improvement, i.e. more calibration samples and/or additional variables should be included.

The last case is not necessarily a problem. It may be a quite interpretable outcome, especially in a one-class problem. A typical example is product quality prediction, which can be done by modeling the single class of acceptable products. If a new sample belongs to the modeled class, it is accepted; otherwise, it is rejected.

Classification And Regression

SIMCA classification can also be based on the X-part of a regression model; read more in the first section hereafter.

Besides, classification may be achieved with a regression technique called Linear Discriminant Analysis, which is an alternative to SIMCA. Read more about the special case PLS Discriminant Analysis in the second section hereafter.

Classification Based on a Regression Model

Throughout this chapter, we have described SIMCA classification as a method involving disjoint PCA modeling. Instead of PCA models, you can also use PCR or PLS models. In those cases, only the X-part of the model will be used. The results will be interpreted in exactly the same way.

SIMCA classification based on the X-part of a regression model is a nice way to detect whether new samples are ***suitable for prediction***. If the samples are recognized as members of the class formed by the calibration sample set, the predictions for those samples should be reliable. Conversely, you should avoid using your model for extrapolation, i.e. making predictions on samples which are rejected by the classification.

PLS Discriminant Analysis

The discriminant analysis approach differs from the SIMCA approach in that it assumes that a sample has to be a member of one of the classes included in the analysis. The most common case is that of a binary discriminant variable: a question with a Yes / No answer.

Binary discriminant analysis is performed using regression, with the ***discriminant variable*** coded 0 / 1 (***Yes = 1, No = 0***) as Y-variable in the model.

With PLS2, this can easily be extended to the case of more than two classes. Each class is represented by an indicator variable, i.e. a binary variable with value 1 for members of that class, 0 for non-members. By building a PLS2 model with all indicator variables as Y, you can directly predict class membership from the X-variables describing the samples. The model is interpreted by viewing **Predicted vs. Measured** for each class indicator Y-variable:

- $Y_{pred} > 0.5$ means "roughly 1" that is to say "***member***";
- $Y_{pred} < 0.5$ means "roughly 0" that is to say "***non-member***".

Once the PLS2 model has been checked and validated (see the chapter about Multivariate Regression p. 107 for more details on diagnosing and validating a model), you can run a Prediction in order to classify new samples.

Interpret the prediction results by viewing the plot **Predicted with Deviations** for each class indicator Y-variable:

- Samples with $Y_{pred} > 0.5$ and a deviation that does not cross the 0.5 line are predicted members;
- Samples with $Y_{pred} < 0.5$ and a deviation that does not cross the 0.5 line are predicted non-members;
- Samples with a deviation that crosses the 0.5 line cannot be safely classified.

See Chapter "Make Predictions" p. 133 for more details on Predicted with Deviations and how to run a prediction.

Classification in Practice

The sections that follow list menu options, dialogs and plots for classification. For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo's web site www.camo.com/TheUnscrambler/Appendices .

Run A Classification

When your data table is displayed in the ***Editor***, you may access the **Task** menu to run a Classification.

Prior to the actual classification, we recommend that you do two things:

1. Insert or append a **category variable** in your data table. This category variable should have as many ***levels*** as you have ***classes***. The easiest way to do this is to define one ***sample set*** for each class, then build the category variable based on the sample sets (this is an option in the Category Variable Wizard).
   The category variable will allow you to use **sample grouping** on PCA and Classification plots, so that each class appears with a different color.
2. Run a **PCA on the training samples** (i.e. the samples with known class membership on which you are going to base the classification model). Check on the score plots for the first PCs (1 vs. 2, 3 vs. 4, 1 vs. 3 etc) whether the classes have a ***good spontaneous separation***. Look for ***outliers*** using warnings, score plots and influence plots. If the classes are not well separated, a ***transformation*** of some variables may be necessary before you can try a classification.

Then the classification procedure itself begins by building ***one PCA model for each class***, diagnosing the models and deciding how many PCs are necessary according to the variance curve (use a proper validation method).

Once all your class PCA models are ***saved***, you may run **Task - Classify**.

Prepare your Data Table for Classification

- **Modify - Edit Set**: Create new sample sets (one for each class + one for all training samples)
- **Edit - Insert - Category Variable**: Insert category variable anywhere in the table
- **Edit - Append - Category Variable**: Add category variable at the right end of the table

Run a global PCA and Check Class Separation

- **Task - PCA**: Run a PCA on all training samples
- **Edit - Options**: Use sample grouping on a score plot

Run Class PCA(s) and Save PCA Model(s)

- **File - Save**: Save PCA model file for the first time, or with existing name
- **File - Save As**: Save PCA model file under a new name

Run Classification

- **Task - Classify**: Run a classification on all training samples
- Later, you may also run a classification on new samples (once you have checked that the training samples are correctly classified)

Save And Retrieve Classification Results

Once the classification has been computed according to your specifications, you may either View the results right away, or Close (and Save) your classification result file to be opened later in the ***Viewer***.

Save Result File from the Viewer

- **File - Save**: Save result file for the first time, or with existing name
- **File - Save As**: Save result file under a new name

Open Result File into a new Viewer

- **File - Open**: Open any file or just lookup file information
- **Results - Classification**: Open classification result file or just lookup file information and warnings
- **Results - All**: Open any result file or just lookup file information, warnings and variances

View Classification Results

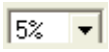
Display classification results as plots from the ***Viewer***. Your classification results file should be opened in the Viewer; you may then access the **Plot** menu to select the various results you want to plot and interpret.

From the **View**, **Edit** and **Window** menus you may use more options to enhance your plots and ease result interpretation.

How To Plot Classification Results

- **Plot - Classification**: Display the classification plots of your choice

More Plotting Options

- **Edit - Options**: Format your plot – on the Sample Grouping sheet, group according to the levels of a category variable
- The 

tool: Change the significance level
- **Edit - Insert Draw Item**: Draw a line or add text to your plot
- **View - Outlier List**: Display list of outlier warnings issued during the analysis
- **Window - Warning List**: Display general warnings issued during the analysis

How To Keep Track of Interesting Objects

- **Edit - Mark**: Several options for marking samples or variables

Run A PLS Discriminant Analysis

When your data table is displayed in the ***Editor***, you may access the **Task** menu to run a Regression (and later on a Prediction).

In order to run a PLS discriminant analysis, you should first prepare your data table in the following way:

1. Insert or append a **category variable** in your data table. This category variable should have as many ***levels*** as you have ***classes***. The easiest way to do this is to define one ***sample set*** for each class, then build the category variable based on the sample sets (this is an option in the Category Variable Wizard).
   The category variable will allow you to use **sample grouping** on PCA and Classification plots, so that each class appears with a different color.
2. **Split** the category variable into indicator variables. These will be your ***Y-variables*** in the PLS model. Create a new ***variable set*** containing only the indicator variables.

Prepare your Data Table for PLS Discriminant Analysis

- **Modify - Edit Set**: Create new sample sets (one for each class + one for all training samples)
- **Edit - Insert - Category Variable**: Insert category variable anywhere in the table
- **Edit - Append - Category Variable**: Add category variable at the right end of the table
- **Edit - Split Category Variable**: Split the category variable into indicator variables
- **Modify - Edit Set**: Create a new variable set (with all indicator variables)

Run a Regression

- **Task - Regression**: Run a regression on all training samples; select PLS as regression method

More options for saving, viewing and refining regression results can be found in chapter "**Multivariate Regression in Practice"** p. 116.

Run a Prediction

- **Task - Predict**: Run a prediction on new samples contained in the current data table

More options for saving and viewing prediction results can be found in chapter "**Prediction in Practice"** p. 135.

Clustering

Use the K-Means algorithm to identify a chosen number of clusters among your samples.

Principles of Clustering

K-Means methodology is a commonly used clustering technique. In this analysis the user starts with a collection of samples and attempts to group them into *'k' Number of Clusters* based on certain specific distance measurements. The prominent steps involved in the K-Means clustering algorithm are given below.

1. This algorithm is initiated by creating 'k' different clusters. The given sample set is first randomly distributed between these 'k' different clusters.
2. As a next step, the distance measurement between each of the sample, within a given cluster, to their respective cluster centroid is calculated.
3. Samples are then moved to a cluster (k′) that records the shortest distance from a sample to the cluster (k′) centroid.

As a first step to the cluster analysis the user decides on the *Number of Clusters 'k'*. This parameter could take definite integer values with the lower bound of 1 (in practice, 2 is the smallest relevant number of clusters) and an upper bound that equals the total number of samples.

The K-Means algorithm is repeated a number of times to obtain an optimal clustering solution, every time starting with a random set of initial clusters.

Distance Types

The following distance types can be used for clustering.

Euclidean distance

This is the most usual, "natural" and intuitive way of computing a distance between two samples. It takes into account the difference between two samples directly, based on the magnitude of changes in the sample levels. This distance type is usually used for data sets that are suitably normalized or without any special distribution problem.

Manhattan distance

Also known as city-block distance, this distance measurement is especially relevant for discrete data sets. While the Euclidean distance corresponds to the length of the shortest path between two samples (i.e. "as the crow flies"), the Manhattan distance refers to the sum of distances along each dimension (i.e. "walking round the block").

Pearson Correlation distance

This distance is based on the Pearson correlation coefficient that is calculated from the sample values and their standard deviations. The correlation coefficient r takes values from –1 (large, negative correlation) to +1 (large, positive correlation).
Effectively, the Pearson distance dp is computed as

$$dp = 1 - r$$

and lies between 0 (when correlation coefficient is +1, i.e. the two samples are most similar) and 2 (when correlation coefficient is -1).
Note that the data are centered by subtracting the mean, and scaled by dividing by the standard deviation.

Absolute Pearson Correlation distance

In this distance, the absolute value of the Pearson correlation coefficient is used; hence the corresponding distance lies between 0 and 1, just like the correlation coefficient.
The equation for the Absolute Pearson distance da is

$$da = 1 - |r|$$

Taking the absolute value gives equal meaning to positive and negative correlations, due to which anti-correlated samples will get clustered together.

Un-centered Correlation distance

This is the same as the Pearson correlation, except that the sample means are set to zero in the expression for un-centered correlation. The un-centered correlation coefficient lies between –1 and +1; hence the distance lies between 0 and 2.

Absolute, Un-centered Correlation distance

This is the same as the Absolute Pearson correlation, except that the sample means are set to zero in the expression for un-centered correlation. The un-centered correlation coefficient lies between 0 and +1; hence the distance lies between 0 and 1.

Kendall's (tau) distance

This non-parametric distance measurement is more useful in identifying samples with a huge deviation in a given data set.

Quality of the Clustering

The clustering analysis results in the assignment of cluster-id to each of the sample based on *Sum Of Distances "SOD"*. The *Sum Of Distances* is described as the sum of the distance values between each of the sample to their respective cluster centroid summed up over all 'k' clusters. This parameter is uniquely calculated and displayed for a particular batch of cluster-ids resulting from a cluster calculation. The results from various different cluster analyses are compared based on the *Sum Of Distances* values. The solution with a least *Sum of Distances* is a good indicator for an acceptable cluster assignment. Hence it is recommended to initiate the analysis with a small *Iteration Number*, say for example 10 for a sample set of 500, and proceed towards a higher cycle of *Iteration Number* to obtain an optimal cluster solution. Once the user obtains an optimal (lowest) *Sum Of Distances* there is a good possibility that there will not be further decline in the *Sum Of Distances* by setting *Iteration Number* to higher values. The cluster-id assignment for an optimal *Sum Of Distances* is considered to be the most appropriate result.

Note: Since the first step of the K-Means algorithm is based on the random distribution of the samples into 'k' different clusters there is a good possibility that the final clustering solution will not be exactly the same for every instance for a fairly large sample data set.

Main Results of Clustering

A clustering analysis gives you the results in form of a ***category variable*** inserted at the beginning of your data table. This category variable has one level (1, 2, …) for each cluster, and tells you which cluster each sample belongs to.

The name of the clustering variable reflects which distance type was applied and how large the SOD was for the retained solution.

For instance, if the clustering was performed using the Euclidean distance, and the best result (the one now displayed in the data table) after 50 iterations was a sum of distances of 80.7654, the clustering variable is called "Euclidean_SOD 80.7654".

Clustering in Practice

This section describes menu options for clustering.

Run A Clustering

When your data table is displayed in the ***Editor***, you may access the **Task** menu to run a Clustering analysis using **Task - Clustering**.

View Clustering Results

The clustering results are stored as a category variable in your data table. Use this variable for sample grouping in plots (either of raw data or of analysis results).

It is recommended to ***run a PCA*** both before and after performing a clustering:

- ***Before***: check for any natural groupings; the PCA score plots may provide you with a relevant number of clusters.
- ***After***: display the new score plots along various PCs with sample grouping according to the clustering variable. This will help you identify which sample properties play an important role in the clustering.

How To Plot Clustering Results

- **Task - PCA**: Run a PCA on your data
- **Plot - Scores**: Display a score plot
- **Plot - Scores and Loadings**: Display a score plot and the corresponding loading plot
- **Edit - Options**: Format your plot – on the Sample Grouping sheet, group according to the levels of the category variable containing clustering results

Analyze Results from Designed Experiments

Specific Methods for Analyzing Designed Data

Assess the important effects and interactions with Analysis of effects, find an optimum with Response surface analysis. Analyze results from Mixture or D-optimal designs with PLS regression.

Simple Data Checks and Graphical Analysis

Any data analysis should start with **simple data checks**: use descriptive statistics, check variable distributions, detect out-of-range values, etc.

For ***designed data***, this stage is even ***more important*** than ever: you would not want to base your test of the significance of the effects on erroneous data, would you?

The good news is that data checks are even ***easier to perform*** when experimental design has helped you generate your data. The reason for this is twofold:

1. If your design variables have any effect at all, the ***experimental design structure*** should be reflected in some way or other in your response data; graphical analyses and PCA will visualize this structure and help you detect features that ***stick out***.
2. The Unscrambler includes ***automatic features*** that take advantage of the design structure (grouping according to ***levels of design variables*** when computing descriptive statistics or viewing a PCA score plot). When the structure of the design shows in the plots (e.g. as sub-groups in a box-plot, or with different colors on a score plot), it is easy for you to spot any sample or variable with an illogical behavior.

General methods for univariate and multivariate descriptive data analysis have been described in the following chapters:

- Describe One Variable At A Time (descriptive statistics and graphical checks) p. 91
- Describe Many Variables Together (Principal Component Analysis) p. 95

These methods apply both to designed and non-designed data. In addition, the sections that follow introduce more specific methods suitable for the analysis of designed data.

Study Main Effects and Interactions

In principle, designed data can be analyzed using the same techniques as non-designed data, i.e. PCA, PCR, PLS or MLR. In addition, The Unscrambler provides several ***specific methods*** that apply particularly well to data from an orthogonal design (Factorial, Plackett-Burman, Box-Behnken or Central Composite).

Among these traditional methods, **Analysis of Effects** is described in this chapter and **Response Surface Modeling** in the next.

The last chapter focuses on the use of PLS for analyzing results from constrained (non-orthogonal) experiments.

What is Analysis of Effects?

The purpose of this method is to find out which design variables have the largest influence on the response variables you have selected, and how significant this influence is. It especially applies to screening designs.

Analysis of Effects includes the following tools:

- ***ANOVA***;
- ***multiple comparisons*** in the case of more than two levels;
- several methods for ***significance testing***.

ANOVA

Analysis of variance (ANOVA) is based on breaking down the variations of a response into several parts that can be compared to each other for significance testing.

To test the ***significance*** of a given effect, you have to compare the variance of the response accounted for by the effect to the residual variance, which summarizes experimental error. If the "structured" variance (due to the effect) is no larger than the "random" variance (error), the effect can be considered negligible. If it is significantly larger than the error, it is regarded as significant.

In practice, this is achieved through a series of successive computations, with results traditionally displayed as a table. The elements listed hereafter define the columns of the ANOVA table, and there is one row for each source of variation:

1. First, several ***sources of variation*** are defined. For instance, if the purpose of the model is to study the main effects of all design variables, each design variable is a source of variation. Experimental error is also a source of variation;
2. Each source of variation has a limited number of independent ways to cause variation in the data. This number is called number of ***degrees of freedom*** (DF);
3. Response variation associated to a specific source is measured by a ***sum of squares*** (SS);
4. Response variance associated to the same source is then computed by dividing the sum of squares by the number of degrees of freedom. This ratio is called ***mean square*** (MS);
5. Once mean squares have been determined for all sources of variation, ***f-ratios*** associated to every tested effect are computed as the ratio of MS(effect) to MS(error). These ratios, which compare structured variance to residual variance, have a statistical distribution which is used for significance testing. The higher the ratio, the more important the effect;
6. Under the null hypothesis (i.e., that the true value of an effect is zero), the f-ratio has a Fisher distribution. This makes it possible to estimate the probability of getting such a high f-ratio under the null hypothesis. This probability is called ***p-value***; the smaller the p-value, the more likely it is that the observed effect is not due to chance. Usually, an effect is declared significant if p-value<0.05 (significance at the 5% level). Other classical thresholds are 0.01 and 0.001.

The outlined sequence of computations applies to all cases of ANOVA. Those can be the following:

- ***Summary ANOVA***: ANOVA on the global model. The purpose is to test the global significance of the whole model before studying the individual effects.
- ***Linear ANOVA***: Each main effect is studied separately.
- ***Linear with Interactions ANOVA***: Each main effect and each 2-factor interaction is studied separately.

- ***Quadratic ANOVA***: Each main effect, each 2-factor interaction and each quadratic effect is studied separately.

Note1: Quadratic ANOVA is not a part of Analysis of Effects, but it is included in Response Surface Analysis (see the next chapter Make a Response Surface Model).

Note2: The underlying computations of ANOVA are based on MLR (see the chapter about Multivariate Regression). The effects are computed from the regression coefficients, according to the following formula:

Main effect of a variable = 2•(b-coefficient of that variable).

Multiple Comparisons

Multiple comparisons apply whenever a design variable with more than two levels has a significant effect. Their purpose is to determine which levels of the design variable have significantly different response meanvalues.

The Unscrambler uses one of the most well-known procedures for multiple comparisons: ***Tukey's Test***. The levels of the design variable are sorted according to their average response value, and non-significantly different levels are displayed together.

Methods for Significance Testing

Apart from ANOVA, which tests the significance of the various effects included in the model, using only the cube samples, Analysis of Effects also provides several other methods for significance testing. They differ from each other by the way the experimental error is estimated. In The Unscrambler, five different sources of experimental error determine different methods.

Higher Order Interaction Effects (HOIE):

Here the ***residual degrees of freedom*** in the ***cube samples*** are used to estimate the experimental error. This is possible whenever the number of effects in the model is substantially smaller than the number of cube samples (e.g. in full factorials designs). Higher order interactions (i.e. interactions involving more than two variables) are assumed to be negligible, thus generating the necessary degrees of freedom. This is the most common method for significance testing, and it is used in the ANOVA computations.

Center samples:

When HOIE cannot be used because of ***insufficient degrees of freedom*** in the cube samples, the experimental error can be estimated from ***replicated center samples***. This is why including several center samples is so useful, especially in fractional factorial designs.

Reference samples:

This method is similar to "center samples", and applies when there are no replicated center samples but some ***reference samples have been replicated***.

Reference and center samples:

When both center and reference samples have been replicated, ***all replicates*** are taken into account to estimate the experimental error.

Comparison with a Scale-Independent Distribution (COSCIND):

If there are ***not enough degrees of freedom*** in the ***cube samples*** and ***no other samples have been replicated***, one degree of freedom can be created by removing the smallest observed effect. Afterwards, the remaining

effects are sorted on increasing absolute value and their significance is estimated using an approximation (the ***Psi statistics***) which is not based on the Fisher distribution. This method has an essentially different philosophy from the others; the p-values computed from the Psi statistic have no absolute meaning. They can only be interpreted in the context of the sorted effects. Going from the smallest effect to the largest, p-value is compared to a significance threshold (e.g. 0.05); when the first significant effect is encountered, all the larger effects can be interpreted as at least as significant.

Whenever such computations are possible, The Unscrambler automatically computes all results based on those five methods. The most relevant one, depending on the context, is then selected as default when you view the results using **Effects Overview**. You can view the results from the other methods if you wish, by selecting another method manually.

Note: When the design includes ***variables with more than two levels***, only HOIE is used.

Make a Response Surface Model

The purpose of Response Surface modeling is to model a response surface using Multiple Linear Regression (MLR). The model can be either linear, linear with interactions, or quadratic. The validity of the model is assessed with the help of ANOVA. The modeled surface can then be plotted to make final interpretation of the results easier.

Read more about MLR in the chapter about Multivariate Regression p. 109.

How to Choose a Response Surface Model

Screening designs, by definition, study only main effects, and possibly interactions. You can use response surface modeling with a **linear** model (with or without **interactions**) to get a 2- or 3-dimensional plot of the effects of two design variables on your responses.

If you wish to analyze results from an ***optimization*** design, the logical choice is a **quadratic** model. This will enable you to check the significance of all effects (linear, interactions, square effects), and to interpret those results (for instance, find the optimum) with the help of the 2- or 3-dimensional plots.

Response Surface Results

Response surface results include the following:

- Leverages;
- Predicted response values;
- Residuals;
- Regression coefficients;
- ANOVA;
- Plots of the response surface.

The first four types of results are classical regression results; lookup Chapter Main Results of Regression p. 111 for more details.

ANOVA and plots include specific features, listed in the sections hereafter.

ANOVA for Linear Response Surfaces

The ANOVA table for a linear response surface includes a few additional features compared to the ANOVA table for analysis of effects (see section ANOVA).

Two new columns are included into the main section showing the individual effects:

- ***b-coefficients***: The values of the regression coefficients are displayed for each effect of the model.
- ***Standard Error of the b-coefficients***: Each regression coefficient is estimated with a certain precision, measured as a standard error.

The Summary ANOVA table also has a new section:

- ***Lack of Fit***: Whenever possible, the error part is divided into two sources of variation, "pure error" and "lack of fit". Pure error is estimated from ***replicated samples***; lack of fit is what remains of the residual sum of squares once pure error has been removed.
  By computing an f-ratio defined by MS(lack of fit)/MS(pure error), the significance of the lack of fit of the model can be tested.
  A significant lack of fit means that the shape of the model does not describe the data adequately. For instance, this can be the case if a linear model is used when there is an important ***curvature***.

ANOVA for Quadratic Response Surfaces

In addition to the above described features, the ANOVA table for a quadratic response surface includes one new column and one new section:

- ***Min/Max/Saddle***: Since the purpose of a quadratic model often is to find out where the optimum is, the minimum or maximum value inside the experimental range is computed, and the design variable values that produce this extreme are displayed as an additional column for the rows where linear effects are tested. Sometimes the extreme is a minimum in one direction of the surface, and a maximum in another direction; such a point is called a ***saddle point***, and it is listed in the same column.
- ***Model Check***: This new section of the table checks the significance of the linear (main effects only) and quadratic (interactions and squares) parts of the model. If the quadratic part is not significant, the quadratic model is too sophisticated and you should try a linear model instead, which will describe your surface more economically and efficiently.

For ***linear models with interactions***, the model check (linear only vs. interactions) is included, but not min/max/saddle.

Response Surface Plots

Specific plots enable you to have a look at the actual shape of the response surface. These plots show the response values as a function of two selected design variables, the remaining variables being constant. The function is computed according to the model equation.

There are two ways to plot a response surface:

- ***Landscape plot***: This plot displays the surface in 3 dimensions, allowing you to study its concrete shape. It is the better type of plot for the visualization of interactions or quadratic effects.
- ***Contour plot***: This plot displays the levels of the response variable as lines on a 2-dimensional plot (like a geographical map with altitudes), so that you can easily estimate the response value for any combination of levels of the design variables. This is done by keeping all variables but two at fixed levels, and plotting the contours of the surface for the remaining two variables. The plot is best suited for final interpretation, i.e. to find the optimum, especially when you need to make a compromise between several responses, or to find a stable region.

Analyze Results from Constrained Experiments

In this section, you will learn how to analyze the results from constrained experiments with methods that take into account the specific features of the design.

The method of choice for the analysis of constrained experiments is PLS regression. If you are not familiar with this method, read about it and how it compares to other regression methods in the chapter on Multivariate Regression (see p. 107).

Use of PLS Regression For Constrained Designs

PLS regression is a **projection method** that decomposes variations within the X-space (predictors, e.g. design variables or mixture proportions) and the Y-space (responses to be predicted) along separate sets of **PLS components** (referred to as PCs). For each dimension of the model (i.e. PC1, PC2, etc.), the summary of X is "biased" so that it is as correlated as possible to the summary of Y. This is how the projection process manages to capture the variations in X that can "explain" variations in Y.

A side effect of the projection principle is that PLS not only builds a model of Y=f(X), it also studies the shape of the multidimensional swarm of points formed by the experimental samples with respect to the X-variables. In other words, it describes the distribution of your samples in the X-space.

Thus any constraints present when building a design, will automatically be detected by PLS because of their impact on the sample distribution. A PLS model therefore has the ability to implicitly take into account Multi-Linear Constraints, mixture constraints, or both. Furthermore, the correlation or even the linear relationships introduced among the predictors by these constraints, will not have any negative effects on the performance or interpretability of a PLS model, contrary to what happens with MLR.

Analyzing Mixture Designs with PLS

When you build a PLS model on the results of mixture experiments, here is what happens:

1. The X-data are centered; i.e. further results will be interpreted as ***deviations from an average situation***, which is the overall centroid of the design;
2. The Y-data are also centered, i.e. further results will be interpreted as an increase or decrease compared to the ***average response values***;
3. The mixture constraint is implicitly taken into account in the model; i.e. the regression coefficients can be interpreted as showing the impact of ***variations*** in each ***mixture component*** when the other ingredients compensate with ***equal proportions***.

In other words: the regression coefficients from a PLS model tell you exactly what happens when you move from the overall centroid towards each corner, along the axes of the simplex.

This property is extremely useful for the analysis of screening mixture experiments: it enables you to interpret the regression coefficients quite naturally as the **main effects** of each mixture component.

The mixture constraint has even more complex consequences on a higher degree model necessary for the analysis of optimization mixture experiments. Here again, PLS performs very well, and the mixture response surface plot enables you to interpret the results visually (see Chapter The Mixture Response Surface Plot p.156 for more details).

Analyzing D-optimal Designs with PLS

PLS regression deals with ***badly conditioned*** experimental matrices (i.e. non-orthogonal X-variables) much better than MLR would do. Actually, the larger the condition number, the more PLS outperforms MLR.

Thus PLS regression is the method of choice to analyze the results from D-optimal designs, no matter whether they involve mixture variables or not.

How Significant are the Results?

The classical methods for significance testing described in the Chapter on Analysis of Effects are not available with PLS regression. However, you may still assess the importance of the effects graphically, and in addition if you cross validate your model you can take advantage of Martens' Uncertainty Test.

Visual Assessment of Effect Importance

In general, the importance of the effects can be assessed visually by looking at the size of the regression coefficients. This is an approximate assessment using the following rule of thumb:

- If the regression coefficient for a variable is ***larger than 0.2*** in absolute value, then the effect of that variable is most probably ***important***.
- If the regression coefficient is ***smaller than 0.1*** in absolute value, then the effect is ***negligible***.
- Between 0.1 and 0.2: "gray zone" where no certain conclusion can be drawn.

Note: In order to be able to compare the relative sizes of your regression coefficients, do not forget to standardize all variables (both X and Y)!

Use of Martens' Uncertainty Test

However, The Unscrambler offers you a much easier, safer and more powerful way of detecting the significance of X-variables: **Martens' Uncertainty Test**. Use this feature in the PLS regression dialog; the significant X-variables will automatically be detected. You will be able to mark them automatically on the regression coefficient plot by using the appropriate icon.

References:

- Martens' Uncertainty Test in chapter "Uncertainty Testing with Cross Validation" p. 123
- Plotting Uncertainty Test results and marking significant variables in chapter "View Regression Results" p. 117

Relevant Regression Models

The ***shape*** of your regression model has to be chosen bearing in mind the objective of the experiments and their analysis. Moreover, the choice of a model plays a significant role in determining which points to include in a design; this applies to classical mixture designs as well as D-optimal designs.

Therefore, The Unscrambler asks you to choose a model immediately after you have defined your design variables, prior to determining the type of classical mixture design or the selection of points building up the D-optimal design which best fits your current purposes.

The minimum number of experiments also depends on the shape of your model; read more about it in Chapter "How Many Experiments Are Necessary?" p. 51.

Models for Non-mixture situations

For constrained designs that do not involve any mixture variables, the choice of a model is straightforward.

Screening designs are based on a **linear** model, with or without interactions. The interactions to be included can be selected freely among all possible products of two design variables.

Optimization designs require a **quadratic** model, which consists of linear terms (main effects), interaction effects, and square terms making it possible to study the curvature of the response surface.

Models for Mixture Variables

As soon of your design involves mixture variables, the ***mixture constraint*** has a remarkable impact on the possible shapes of your model. Since the sum of the mixture components is constant, each mixture component can be expressed as a function of the others. As a consequence, the terms of the model are also linked and you are not free to select any combination of linear, interaction or quadratic terms you may fancy.

Note: In a mixture design, the ***interaction*** and ***square*** effects are ***linked*** and cannot be studied separately.

Example:

A, B and C vary from 0 to 1.

A+B+C = 1 for all mixtures.

Therefore, **C** can be re-written as **1 - (A+B)**.

As a consequence, the square effect C*C or $\mathbf{C^2}$ can also be re-written as $(1-(A+B))^2 = \mathbf{1 + A^2 + B^2 -2A - 2B + 2A*B}$:

it does not make any sense to try to interpret square effects independently from main effects and interactions.

In the same way, **A*C** can be re-expressed as A*(1-A-B) = **A - A*A - A*B**, which shows that interactions cannot be interpreted without also taking into account main effects and square effects.

Here are therefore the basic principles for building relevant mixture models:

Mixture Models for Screening

For **screening** purposes, use a **purely linear** model (without any interactions) with respect to the mixture components.

Important! If your design includes process variables, their interactions with the mixture components may be included, provided that each process variable is combined with either all or none of the mixture variables.

That is to say that if you include the interaction between a process variable P and a mixture variable M1 (interaction PxM1), you must also include interactions PxM2, PxM3,… between this same process variable and all of the other mixture variables. No restriction is placed on the interactions among the process variables themselves.

Make a model with the right selection of variables and interactions in the Regression dialog; or after a first model by marking them on the regression coefficients plot and using **Task - Recalculate with Marked**.

Mixture Models for Optimization

For **optimization** purposes, you will choose a **full quadratic** model with respect to the mixture components.

If any process variables are included in the design, their square effects may or may not be studied, independently of their interactions and of the shape of the mixture part of the model. But as soon as you are interested in process-mixture interactions, the same restriction as before applies.

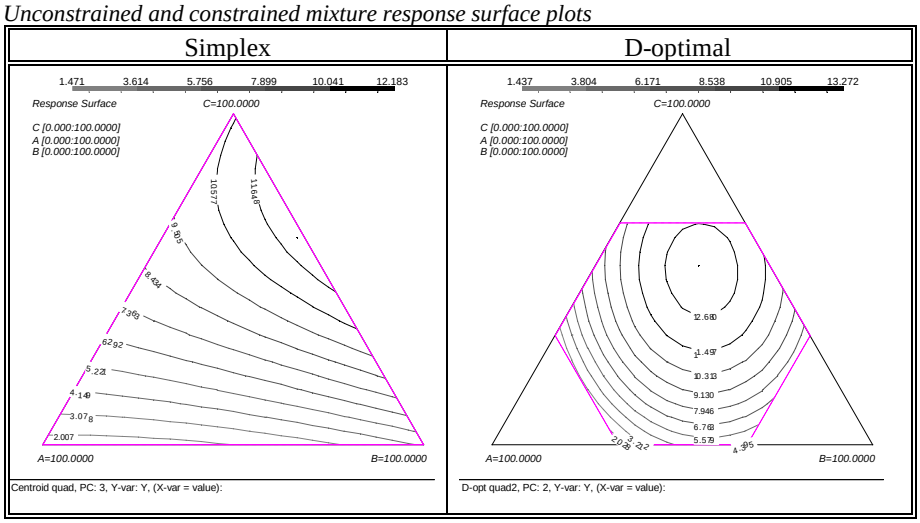
The Mixture Response Surface Plot

Since the mixture components are linked by the ***mixture constraint***, and the experimental region is based on a simplex, a mixture response surface plot has a ***special shape*** and is computed according to special rules.

Instead of having two coordinates, the mixture response surface plot uses a ***special system of 3 coordinates***. Two of the coordinate variables are varied independently from each other (within the allowed limits of course), and the third one is computed as the difference between MixSum and the other two.

Examples of mixture response surface plots, with or without additional constraints, are shown in the figure below.

Unconstrained and constrained mixture response surface plots



Similar response surface plots can also be built when the design includes one or several process variables.

Analyzing Designed Data in Practice

The sections that follow list menu options, dialogs and plots for the analysis of designed data. For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo's web site www.camo.com/TheUnscrambler/Appendices .

Run an Analysis on Designed Data

When your data table is displayed in the ***Editor***, you may access the **Task** menu to run a suitable analysis.

- **Task - Statistics**: Compute Descriptive Statistics on the current data table
- **Task - PCA**: Run a PCA on the current data table
- **Task - Analysis of Effects**: Run an Analysis of Effects on the current data table
- **Task - Response Surface**: Run a Response Surface analysis on the current data table
- **Task - Regression**: Run a regression on the current data table (choose method PLS for constrained designs)

Save And Retrieve Your Results

Once the analysis has been performed according to your specifications, you may either View the results right away, or Close (and Save) your result file to be opened later in the ***Viewer***.

Save Result File from the Viewer

- **File - Save**: Save result file for the first time, or with existing name
- **File - Save As**: Save result file under a new name

Open Result File into a new Viewer

- **File - Open**: Open any file or just lookup file information
- **Results - PCA**, **Results - Statistics**, etc.: Open a specific type of result file or just lookup file information, warnings and variances
- **Results - All**: Open any result file or just lookup file information, warnings and variances

Display Data Plots and Descriptive Statistics

This topic is fully covered in Chapter "Univariate Data Analysis in Practice" p. 92.

View Analysis of Effects Results

Display Analysis of Effects results as plots from the ***Viewer***. Your results file should be opened in the Viewer; you may then access the **Plot** menu to select the various results you want to plot and interpret.

From the **View**, **Edit** and **Window** menus you may use more options to enhance your plots and ease result interpretation.

How To Plot Analysis of Effects Results

- **Plot - Effects**: Display the main plot of effects (and select appropriate significance testing method)
- **Plot - Analysis of Variance**: Display ANOVA table
- **Plot - Residuals**: Display various types of residual plots
- **Plot - Predicted vs Measured**: Display plot of predicted Y values against actual Y values
- **Plot - Response Surface**: Plot predicted Y values as a function of 2 design variables

PC Navigation Tool

Navigate up or down the PCs in your model along the vertical and horizontal axes of your plots:

- **View - Source - Previous Vertical PC**
- **View - Source - Next Vertical PC**
- **View - Source - Back to Suggested PC**
- **View - Source - Previous Horizontal PC**
- **View - Source - Next Horizontal PC**

More Plotting Options

- **Edit - Options**: Format your plot
- **Edit - Insert Draw Item**: Draw a line or add text to your plot

- **View - Outlier List**: Display list of outlier warnings issued during the analysis for each PC, sample and/or variable
- **Window - Warning List**: Display general warnings issued during the analysis

How To Change Plot Ranges:

- **View - Scaling**
- **View - Zoom In**
- **View - Zoom Out**

How To Keep Track of Interesting Objects

- **Edit - Mark**: Several options for marking samples or variables

View Response Surface Results

Display response surface results as plots from the ***Viewer***. Your results file should be opened in the Viewer; you may then access the **Plot** menu to select the various results you want to plot and interpret.

From the **View**, **Edit** and **Window** menus you may use more options to enhance your plots and ease result interpretation.

How To Plot Response Surface Results

- **Plot - Response Surface Overview**: Display the 4 main response surface plots
- **Plot - Response Surface**: Display the a response surface plot according to your specifications
- **Plot - Analysis of Variance**: Display ANOVA table (MLR)
- **Plot - Residuals**: Display various types of residual plots
- **Plot - Regression Coefficients**: Plot regression coefficients
- **Plot - Predicted vs Measured**: Display plot of predicted Y values against actual Y values
- **Plot - Regression and Prediction**: Display Predicted vs. Measured and Regression coefficients
- **Plot - Leverage**: Plot sample leverages

More Plotting Options

- **Edit - Options**: Format your plot
- **Edit - Insert Draw Item**: Draw a line or add text to your plot
- **View - Outlier List**: Display list of outlier warnings issued during the analysis for each PC, sample and/or variable
- **Window - Warning List**: Display general warnings issued during the analysis
- **View - Toolbars**: Select which groups of tools to display on the toolbar
- **Window - Identification**: Display curve information for the current plot

How To Change Plot Ranges:

- **View - Scaling**

- **View - Zoom In**
- **View - Zoom Out**

How To Keep Track of Interesting Objects

- **Edit - Mark**: Several options for marking samples or variables

View Regression Results for Designed Data

This topic is fully covered in Chapter “View Regression Results” p. 117.

Multivariate Curve Resolution

The theoretical sections of this chapter were authored by Romà Tauler and Anna de Juan.

Principles of Multivariate Curve Resolution (MCR)

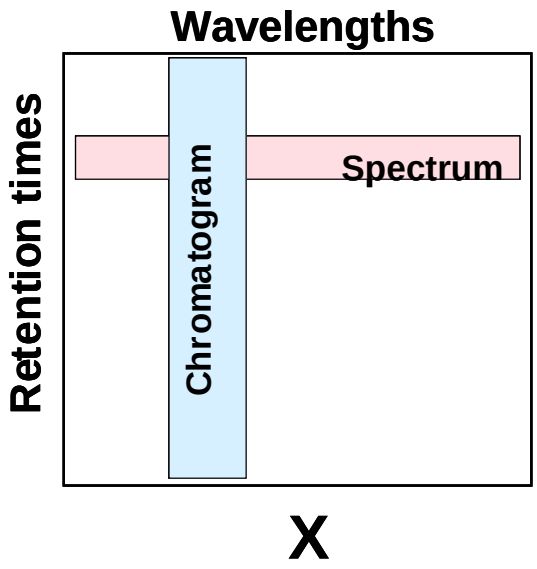
Most of the data examples analyzed until now were arranged in two-way data "flat" table structures. An alternative to PCA in the analysis of these two-way data tables is to perform MCR on them.

What is MCR?

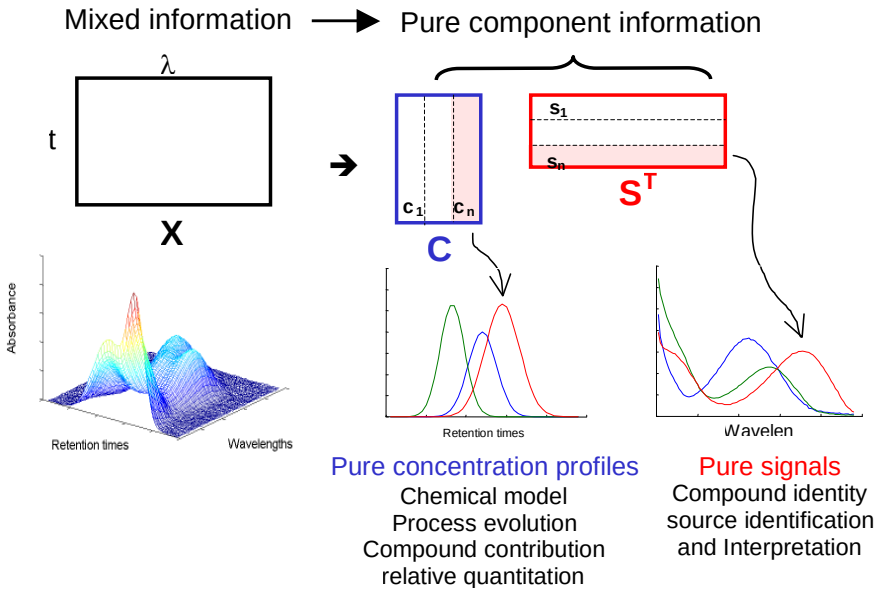
Multivariate Curve Resolution (MCR) methods may be defined as a group of techniques which intend the recovery of concentration (pH profiles, time/kinetic profiles, elution profiles, chemical composition changes...) and response profiles (spectra, voltammograms...) of the components in an unresolved mixture using a minimal number of assumptions about the nature and composition of these mixtures. MCR methods can be easily extended to the analysis of many types of experimental data including multi-way data.

Data Suitable for MCR

A typical example is related to chromatographic hyphenated techniques, like liquid chromatography with diode array detection (LC-DAD), where a set of UV-VIS spectra are obtained at the different elution times of the chromatographic run. Then, the data may be arranged in a data table, where the different spectra at the different elution times are set in the rows and the elution profiles changing with time at the different wavelengths are set in the columns. So, in the analysis of a single sample, a table or data matrix **X** is obtained:



Multivariate Curve Resolution (MCR)



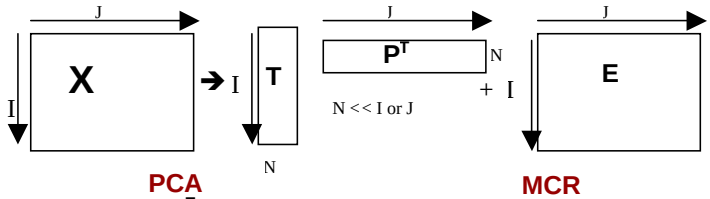
Pure concentration profiles
Chemical model
Process evolution
Compound contribution
relative quantitation

Pure signals
Compound identity
source identification
and Interpretation

Purposes of MCR

Multivariate Curve Resolution has been shown to be a powerful tool to describe multi-component mixture systems through a **bilinear** model of **pure component contributions**. MCR, like PCA, assumes the fulfilment of a bilinear model, i.e

Bilinear models for two-way data



PCA
T orthogonal, $\mathbf{P}^T$ orthonormal
$\mathbf{P}^T$ in the direction of maximum variance

Unique solutions
but without physical meaning
Useful for interpretation

MCR
Other constraints (non-negativity, unimodality, local rank,…)
T=C and $\mathbf{P}^T = \mathbf{S}^T$ non-negative,...
C or $\mathbf{S}^T$ normalization

Non-unique solutions
but with physical meaning
Useful for resolution (and obviously for interpretation)!

Limitations of PCA

Principal Component Analysis, PCA, produces an ***orthogonal*** bilinear matrix decomposition, where components or factors are obtained in a sequential way explaining maximum variance. Using these constraints plus normalization during the bilinear matrix decomposition, PCA produces ***unique*** solutions. These 'abstract' unique and orthogonal (independent) solutions are very helpful in deducing the number of different sources of variation present in the data and, eventually, they allow for their identification and interpretation. However, these solutions are 'abstract' solutions in the sense that they are not the 'true' underlying factors causing the data variation, but orthogonal linear combinations of them.

The Alternative: Curve Resolution

On the other hand, in Curve Resolution methods, the goal is to unravel the ***'true' underlying sources of data variation***. It is not only a question of how many different sources are present and how they can be interpreted, but to find out how they are in reality. The price to pay is that unique solutions are not usually obtained by means of Curve Resolution methods unless external information is provided during the matrix decomposition.

Whenever the goals of Curve Resolution are achieved, the understanding of a chemical system is dramatically increased and facilitated, avoiding the use of enhanced and much more costly experimental techniques. Through Multivariate Resolution methods, the ubiquitous mixture analysis problem in Chemistry (and other scientific fields) is solved directly by mathematical and software tools instead of using costly analytical chemistry and instrumental tools, for example as in sophisticated hyphenated mass spectrometry-chromatographic methods.

The next sections will present the following topics:

- How unique is the MCR solution? in “Rotational and Intensity Ambiguities in MCR” p. 165
- How to take into account additional information: “Constraints in MCR” p.165
- MCR results in “Main Results of MCR” p.163
- Types of problems which MCR can solve in “MCR Application Examples” p.168

As a comparison, you may also read more about PCA in chapter “Principles of Projection and PCA” p. 95.

You may also read about the MCR-ALS algorithm in the Method Reference chapter, available as a separate .PDF document for easy print-out of the algorithms and formulas – download it from Camo’s web site www.camo.com/TheUnscrambler/Appendices.

Main Results of MCR

Contrary to what happens when you build a PCA model, the number of components computed in MCR is not your choice. The optimal number of components **n** necessary to resolve the data is estimated by the system, and the total number of components saved in the MCR model is set to **n+1**.

Note: As there must be at least two components in a mixture, the minimum number of components in MCR is 2.

For each number of components **k** between **2** and **n+1**, the MCR results are as follows:

- ***Residuals*** are error measures; they tell you how much variation remains in the data after **k** components have been estimated;
- ***Estimated concentrations*** describe the estimated pure components’ profiles across all the samples included in the model;
- ***Estimated spectra*** describe the instrumental properties (e.g. spectra) of the estimated pure components.

Residuals

The residuals are a measure of the fit (or rather, misfit) of the model. The smaller the residuals, the better the fit.

MCR residuals can be studied from three different points of view.

- ***Variable Residuals*** are a measure of the variation remaining in each variable after **k** components have been estimated. In The Unscrambler, the variable residuals are plotted as a **line plot** where each variable is represented by one value: its residual in the **k**-component model.
- ***Sample Residuals*** are a measure of the distance between each sample and its model approximation. In The Unscrambler, the sample residuals are plotted as a **line plot** where each sample is represented by one value: its residual after **k** components have been estimated.
- ***Total Residuals*** express how much variation in the data ***remains to be explained*** after **k** components have been estimated. Their role in the interpretation of MCR results is similar to that of Variances in PCA. They are plotted as a **line plot** showing the total residual after a varying number of components (from **2** to **n+1**).

The three types of MCR residuals are available for two different model fits.

- **MCR Fitting**: these are the actual values of the residuals after the data have been resolved to **k** pure components.
- **PCA Fitting**: these are the residuals from a PCA with **k** PCs performed on the same data.

Estimated Concentrations

The estimated concentrations show the profile of each estimated pure component across the samples included in the MCR model.

In The Unscrambler, the estimated concentrations are plotted as a **line plot** where the abscissa shows the samples, and each of the **k** pure components is represented by one curve.

The **k** estimated concentration profiles can be interpreted as **k** new variables telling you how much each of your original samples contains of each estimated pure component.

Note!

Estimated concentrations are expressed as relative values within individual components. The estimated concentrations for a sample are ***not*** its real composition.

Estimated Spectra

The estimated spectra show the estimated instrumental profile (e.g. spectrum) of each pure component across the X-variables included in the analysis.

In The Unscrambler, the estimated spectra are plotted as a **line plot** where the abscissa shows the X-variables, and each of the **k** pure components is represented by one curve.

The **k** estimated spectra can be interpreted as the spectra of **k** new samples consisting each of the pure components estimated by the model. You may compare the spectra of your original samples to the estimated spectra so as to find out which of your actual samples are closest to the pure components.

Note!

Estimated spectra are unit-vector normalized.

More Details About MCR

Rotational and Intensity Ambiguities in MCR

From the early days in resolution research, the mathematical decomposition of a single data matrix, no matter the method used, has been known to be subject to ambiguities. This means that many pairs of **C**- and $\mathbf{S}^T$-type matrices can be found that reproduce the original data set with the same fit quality. In plain words, the correct reproduction of the original data matrix can be achieved by using component profiles differing in shape (rotational ambiguity) or in magnitude (intensity ambiguity) from the sought (true) ones.

These two kinds of ambiguities can be easily explained. The basic equation associated with resolution methods, $\mathbf{X} = \mathbf{C}\,\mathbf{S}^T$, can be transformed as follows:

$$\mathbf{X} = \mathbf{C}\,(\mathbf{T}\,\mathbf{T}^{-1})\,\mathbf{S}^T$$
$$\mathbf{X} = (\mathbf{C}\,\mathbf{T})\,(\mathbf{T}^{-1}\,\mathbf{S}^T)$$
$$\mathbf{X} = \mathbf{C'}\,\mathbf{S'}^T$$

where $\mathbf{C'} = \mathbf{C}\,\mathbf{T}$ and $\mathbf{S'}^T = (\mathbf{T}^{-1}\,\mathbf{S}^T)$ describe the **X** matrix as correctly as the true **C** and $\mathbf{S}^T$ matrices do, though **C'** and $\mathbf{S'}^T$ are not the sought solutions. As a result of the rotational ambiguity problem, a resolution method can potentially provide as many solutions as **T** matrices can exist. This may represent an infinite set of solutions, unless **C** and $\mathbf{S}^T$ are forced to obey certain conditions. In a hypothetical case with no rotational ambiguity, that is, the shapes of the profiles in **C** and $\mathbf{S}^T$ are correctly recovered, the basic resolution model with intensity ambiguity could be written as shown below:

$$\mathbf{X} = \sum_{i=1}^{n} \left(\frac{1}{k_i} \mathbf{c}_i \right) \left(k_i \mathbf{s}_i^T \right)$$

where k_i are scalars and n refers to the number of components. Each concentration profile of the new **C'** matrix would have the same shape as the real one, but being k_i times smaller, whereas the related spectra of the new **S'** matrix would be equal in shape to the real spectra, though k_i times more intense.

Constraints in MCR

Although resolution does not require previous information about the chemical system under study, additional knowledge, when it exists, can be used to tailor the sought pure profiles according to certain known features and, as a consequence, to minimize the ambiguity in the data decomposition and in the results obtained.

The introduction of this information is carried out through the implementation of ***constraints***.

What is a Constraint?

A constraint can be defined as any mathematical or chemical property systematically fulfilled by the whole system or by some of its pure contributions. Constraints are translated into mathematical language and force the iterative optimization to model the profiles respecting the conditions desired.

When to apply a Constraint

The application of constraints should be always prudent and soundly grounded and they should only be set when there is an absolute certainty about the validity of the constraint. Even a potentially useful constraint can play a negative role in the resolution process when factors like experimental noise or instrumental problems distort the related profile or when the profile is modified so roughly that the convergence of the optimization process is seriously damaged. When well implemented and fulfilled by the data set, constraints can be seen as the driving forces of the iterative process to the right solution and, often, they are found not to be active in the last part of the optimization process.

The efficient and reliable use of constraints has improved significantly with the development of methods and software that allow them to be easily used in flexible ways. This increase in flexibility allows complete

freedom in the way combinations of constraints may be used for profiles in the different concentration and spectral domains. This increase in flexibility also makes it possible to apply a certain constraint with variable degrees of tolerance to cope with noisy real data, i.e., the implementation of constraints often allows for small deviations from the ideal behavior before correcting a profile. Methods to correct the profile to be constrained have evolved into smoother methodologies, which modify the wrong-behaved profile so that the global shape is kept as much as possible and the convergence of the iterative optimization is minimally upset.

Constraint Types in MCR

There are several ways to classify constraints: the main ones relate either to the nature of the constraints or to the way they are implemented. In terms of their nature, constraints can be based on either chemical or mathematical features of the data set. In terms of implementation, we can distinguish between equality constraints or inequality constraints. An **equality** constraint sets the elements in a profile to be equal to a certain value, whereas an **inequality** constraint forces the elements in a profile to be unequal (higher or lower) than a certain value. The most widely used types of constraints will be described using these classification schemes. In some of the descriptions that follow, comments on the implementation (as equality or inequality constraints) will be added to illustrate this concept.

Non-negativity

The **non-negativity constraint** is applied when it can be assumed that the measured values in an experiment will always be non-negative.

This constraint forces the values in a profile to be equal to or greater than zero. It is an example of an ***inequality*** constraint.

Non-negativity constraints may be applied independently of each other to

- **Concentrations** (the elements in each row of the **C** matrix)
- **Response profiles** (the elements in each row of the $\mathbf{S}^T$ matrix)

For example, non-negativity applies to:

- All concentration profiles in general;

- Many instrumental responses, such as UV absorbances, fluorescence intensities etc.

Unimodality

The **unimodality constraint** allows the presence of only one maximum per profile.

This condition is fulfilled by many peak-shaped concentration profiles, like chromatograms, by some types of reaction profiles and by some instrumental signals, like certain voltammetric responses.

It is important to note that this constraint does not only apply to peaks, but to profiles that have a constant maximum (plateau) and a decreasing tendency. This is the case of many monotonic reaction profiles that show only the decay or the emergence of a compound, such as the most protonated and deprotonated species in an acid-base titration reaction, respectively.

Closure

The **closure constraint** is applied to closed reaction systems, where the principle of mass balance is fulfilled. With this constraint, the sum of the concentrations of all the species involved in the reaction (the suitable elements in each row of the **C** matrix) is forced to be equal to a constant value (the total concentration) at each stage in the reaction. The closure constraint is an example of ***equality*** constraint.

In practice, the closure constraint in MCR forces the sum of the concentrations of all the mixture components to be equal to a constant value (the total concentration) across all samples included in the model.

Other constraints

Apart from the three constraints previously defined, other types of constraints can be applied. See literature on curve resolution for more information about them.

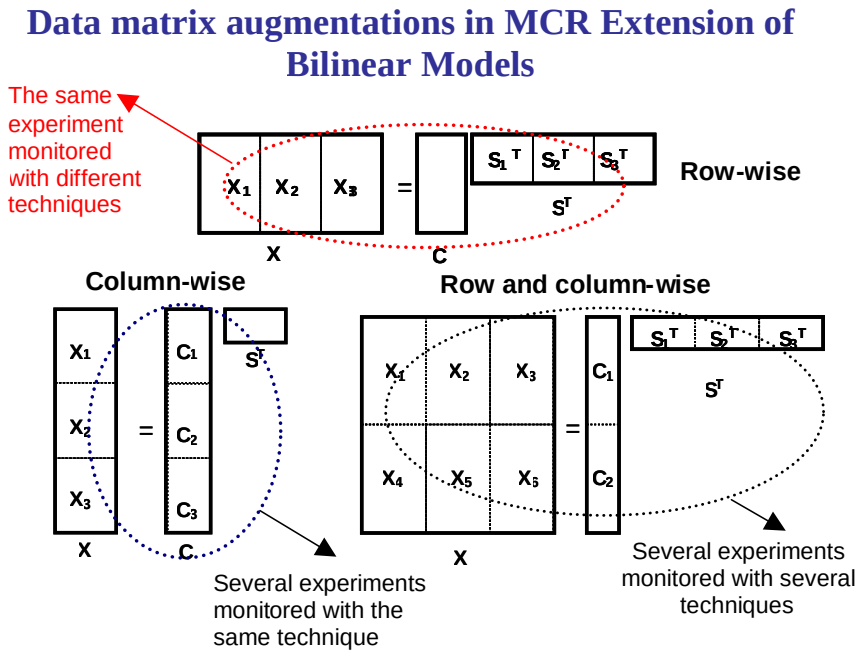
Local rank constraints

Particularly important for the correct resolution of two-way data systems are the so-called *local rank constraints, selectivity and zero-concentration windows*. These types of constraints are associated with the concept of local rank, which describes how the number and distribution of components varies locally along the data set. The key constraint within this family is selectivity. Selectivity constraints can be used in concentration and spectral windows where only one component is present to completely suppress the ambiguity linked to the complementary profile in the system. Thus, selective concentration windows provide unique spectra of the associated components and vice versa. The powerful effect of this type of constraints and their direct link with the corresponding concept of chemical selectivity explains their early and wide application in resolution problems. Not so common, but equally recommended is the use of other local rank constraints in iterative resolution methods. These types of constraints can be used to describe which components are absent in data set windows by setting the number of components inside windows smaller than the total rank. This approach always improves the resolution of profiles and minimizes the rotational ambiguity in the final results.

Physico-chemical constraints

One of the most recent progresses in chemical constraints refers to the implementation of a physicochemical model into the multivariate curve resolution process. In this manner, the concentration profiles of compounds involved in a kinetic or a thermodynamic process are shaped according to the suitable chemical law. Such a strategy has been used to reconcile the separate worlds of hard- and soft-modeling and has enabled the mathematical resolution of chemical systems that could not be successfully tackled by either of these two pure methodologies alone. The strictness of the hard model constraints dramatically decreases the ambiguity of the constrained profiles and provides fitted parameters of physicochemical and analytical interest, such as equilibrium constants, kinetic rate constants and total analyte concentrations. The soft-part of the algorithm allows for modeling of complex systems, where the central reaction system evolves in the presence of absorbing interferences.

Finally, it should be mentioned that MCR methods based on a bilinear model may be easily adapted to resolve three-way data sets. Particular multi-way models and structures may be easily implemented in the form of constraints during MCR optimization algorithms, such as Alternating Least Squares (see below). The discussion of this topic is, however, out of the scope of the present chapter. When a set of data matrices is obtained in the analysis of the same chemical system, they can be simultaneously analyzed setting all of them together in an augmented data matrix and following the same steps as for a single data matrix analysis. The possible data arrangements are displayed in the following figure:


Data matrix augmentations in MCR Extension of Bilinear Models
The same experiment monitored with different techniques
Row-wise
Column-wise
Row and column-wise
Several experiments monitored with the same technique
Several experiments monitored with several techniques

MCR Application Examples

This section briefly presents two application examples.

Note! What follows is not a tutorial. See the Tutorials chapter for more examples and hands-on training.

Solving Co-elution Problems in LC-DAD Data

A classical application of MCR-ALS is the resolution of the co-elution peak of a mixture.

A mixture of three compounds co-elutes in a LC-DAD analysis, i.e. their elution profiles and UV spectra overlap. Spectra are collected at different elution times, and the corresponding chromatograms are measured at the different wavelengths.

First, the number of components can be easily deduced from rank analysis of the data matrix, for instance, using PCA. Then initial estimates of spectra or elution profiles for these three compounds are obtained to start the ALS iterative optimization. Possible constraints to be applied are non-negativity for elution and spectra profiles, unimodality for elution profiles and a type of normalization to scale the solutions. Normalization of spectra profiles may also be recommended.

Reference:

R. Tauler, S. Lacorte and D. Barceló. "Application of multivariate curve self-modeling curve resolution for the quantitation of trace levels of organophosphorous pesticides in natural waters from interlaboratory studies". J. of Chromatogr. A, **730**, 177-183 (1996).

Spectroscopic Monitoring of a Chemical Reaction or Process

A second example frequently encountered in curve resolution studies is the study and analysis of chemical reactions or processes monitored using spectroscopic methods. The process may evolve with time or because some master variable of the system changes, like pH, temperature, concentration of reagents or any other

property. For example in the case of an A → B reaction where both A and B have overlapped spectra, and reaction profiles also overlap in the whole range of study.

This is a case of strong rotational ambiguity since many possible solutions to the problem are possible. Using non-negativity (for both spectra and reaction profiles) unimodality and closure (for reaction profiles) reduces considerably the number of possible solutions.

Alternating Least Squares (MCR-ALS): An Algorithm to Solve MCR Problems

Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) uses an iterative approach to find the matrices of concentration profiles and instrumental responses. In this method, neither **C** nor $\mathbf{S}^T$ matrices have priority over each other and both are optimized at each iterative cycle.

The MCR-ALS algorithm is described in detail in the Method Reference chapter, available as a separate .PDF document for easy print-out of the algorithms and formulas – download it from Camo's web site www.camo.com/TheUnscrambler/Appendices.

Initial estimates for MCR-ALS

Starting the iterative optimization of the profiles in **C** or $\mathbf{S}^T$ requires a matrix or a set of profiles sized as **C** or as $\mathbf{S}^T$ with more or less rough approximations of the concentration profiles or spectra that will be obtained as the final results. This matrix contains the *initial estimates* of the resolution process. In general, the use of ***non-random estimates*** helps shorten the iterative optimization process and helps to avoid convergence to local optima different from the desired solution. It is sensible to use chemically meaningful estimates if we have a way of obtaining them or if the necessary information is available.

Whether the initial estimates are either a **C**-type or an $\mathbf{S}^T$-type matrix can depend on which type of profiles are less overlapped, which direction of the matrix (rows or columns) has more information or simply on the will of the chemist.

In The Unscrambler, you have the possibility to enter your own estimates as **initial guess**.

How To Interpret MCR Results

Once an MCR model is built, you have to ***diagnose*** it, i.e. assess its quality, before you can actually use it for ***interpretation***.

There are two types of factors that may affect the quality of the model:

1. Computational parameters;
2. Quality of the data.

The sections that follow explain what can be done to improve the quality of a model. It may take several improvement steps before you are satisfied with your model.

Once the model is found satisfactory, you may interpret the MCR results and apply them to a better understanding of the system you are studying (e.g. chemical reaction mechanism or process). The last section hereafter will show you how.

Computational Parameters of MCR

In the Unscrambler MCR procedure, the computational parameters for which user input is allowed are the *constraint* settings (non-negative concentrations, non-negative spectra, unimodality, closure) and the setting for *Sensitivity to pure components*.

Read more about:

- When to apply constraints, in chapter "Constraint Settings Are Known Beforehand" below.
- "How To Tune Sensitivity to Pure Components" p.170.

Constraint Settings Are Known Beforehand

In general, you know which constraints apply to your application and your data before you start building the MCR model.

Example (courtesy of Prof. Chris Brown, University of Rhode Island, USA):

FTIR is employed to monitor the reaction of iso-propanol and acetic anhydride using pyridine as a catalyst in a carbon tetrachloride solution. Iso-propyl acetate is one of the products in this typical esterification reaction.

As long as nothing more is added to the samples in the course of the reaction, the sum of the concentrations of the pure components (iso-propanol, acetic anhydride, pyridine, iso-propyl acetate + possibly other products of the esterification) should remain constant. This satisfies the requirements for a ***closure*** constraint.

Of course, if you realize upon viewing your results that the sum of the estimated concentrations is not constant – whereas you know that it *should* be – you can always introduce a closure constraint next time you recalculate the model.

Read more about:

- "Constraints in MCR" p.165

How To Tune Sensitivity to Pure Components

Example: The case of very small components

Unlike the constraints applying to the system under study, which usually are known beforehand, you may have little information about the relative order of magnitude of the estimated pure components upon your first attempt at curve resolution.

For instance, one of the products of the reaction may be ***dominating***, but you are still interested in detecting and identifying ***possible by-products***.

If some of these by-products are synthesized in a very small amount compared to the initial chemicals present in the system and the main product of the reaction, the MCR computations will have trouble distinguishing these by-products' "signature" from mere noise in the data.

General use of Sensitivity to pure components

This is where tuning the parameter called "*sensitivity to pure components*" may help you. This unitless number with formula

Ratio of Eigenvalues E1/(En*10)

can be roughly interpreted as how dominating the last estimated primary principal component is (the one that generates the weakest structure in the data), compared to the first one. The higher the sensitivity, the more pure components will be extracted (the MCR procedure will allow the last component to be more "negligible" in comparison to the first one).

By default, a value of 100 is used; you may tune it up or down between 10 and 190 if necessary.

Read what follows for concrete situation examples.

When to tune Sensitivity up or down

Upon viewing your first MCR results, check the estimated number of pure components and study the profiles of those components.

Case 1: The estimated number of pure components is ***larger than expected***. Action: *reduce sensitivity*.

Case 2: You have no prior expectations about the number of pure components, but some of the extracted profiles look very ***noisy*** and/or ***two of the estimated spectra are very similar***. This indicates that the actual number of components is probably smaller than the estimated number. Action: *reduce sensitivity.*

Case 3: You know that there are at least **n** different components whose concentrations vary in your system, and the estimated number of pure components is ***smaller than*** **n**. Action: *increase sensitivity*.

Case 4: You know that the system should contain a trace-level component, which is ***not detected*** in the current resolution. Action: *increase sensitivity.*

Case 5: You have no prior expectations about the number of pure components, and you are ***not sure*** whether the current results are sensible or not. Action: *check MCR message list.*

Use of the MCR Message List

One of the diagnostic tools available upon viewing MCR results is the MCR Message List, accessed by clicking **View - MCR Message List**. This small box provides you with system recommendations (based on some numerical properties of the results) regarding the value of the MCR parameter Sensitivity to pure components and the possible need for some data pre-processing.

There are four types of recommendations:

- Type 1: Increase sensitivity to pure components;
- Type 2: Decrease sensitivity to pure components;
- Type 3: Change sensitivity to pure components (increase or decrease);
- Type 4: Baseline offset or normalization is recommended.

If none of the above applies, the text "No recommendation" is displayed. Otherwise, you should try the recommended course of action and compare the new results to the old ones.

Outliers in MCR

As in any other multivariate analysis, the available data may be more or less "clean" when you build your first curve resolution model.

The main tool for diagnosing ***outliers*** in MCR consists of two plots of sample residuals, accessed with menu option **Plot - Residuals**.

Any sample that sticks out on the plots of Sample Residuals (either with MCR fitting or PCA fitting) is a possible outlier.

To find out more about such a sample (Why is it outlying? Is it an influential sample? Is that sample dangerous for the model?), it is recommended to run a PCA on your data.

If you find out that the outlier should be removed, you may recalculate the MCR model without that sample.

Read more about:

- "Residuals in MCR" p.164
- "How to detect outliers with PCA" p. 101

Noisy Variables in MCR

In MCR, some of the available variables – even if, strictly speaking, they are no more "noisy" than the others – may contribute poorly to the resolution, or even disturb the results.

The two main cases are:

- ***Non-targeted*** wavelength regions: these variables carry virtually no information that can be of use to the model;
- ***Highly overlapped*** wavelength regions: several of the estimated components have simultaneous peaks in those regions, so that their respective contributions are difficult to entangle.

The main tool for diagnosing ***noisy variables*** in MCR consists of two plots of variable residuals, accessed with menu option **Plot - Residuals**.

Any variable that sticks out on the plots of Variable Residuals (either with MCR fitting or PCA fitting) may be disturbing the model, thus reducing the quality of the resolution; try recalculating the MCR model without that variable.

Practical Use of Estimated Concentrations and Spectra

Once you have managed to build an MCR model that you find satisfactory, it is time to interpret the results and make practical use of the main findings.

The results can be interpreted from three different points of view:

1. Assess or confirm the ***number of pure components*** in the system under study;
2. ***Identify*** the extracted components, using the estimated spectra;
3. ***Quantify*** variations across samples, using the estimated concentrations.

Here are a few rules and principles that may help you:

1. To have reliable results on the *number of pure components*, you should cross-check with a PCA result, try different settings for the Sensitivity to pure components, and use the navigation bar to study the MCR results for various estimated numbers of pure components.

2. *Weak* components (either low concentration or noise) are usually listed first.

3. Estimated spectra are *unit-vector normalized.*

4. The spectral profiles obtained may be compared to a library of similar spectra in order to identify the *nature* of the pure components that were resolved.

5. Estimated concentrations are *relative values* within an individual component itself. Estimated concentrations of a sample are NOT its real composition.

Application examples:

1. One can utilize *estimated concentration profiles* and other experimental information to analyze a chemical/ biochemical reaction mechanism.

2. One can utilize *estimated spectral profiles* to study the mixture composition or even intermediates during a chemical/biochemical process.

Multivariate Curve Resolution in Practice

The sections that follow list menu options, dialogs and plots for multivariate curve resolution. For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo's web site www.camo.com/TheUnscrambler/Appendices.

In practice, building and using an MCR model consists of several steps:

1. Choose and implement an appropriate ***pre-processing*** method (see Chapter Re-formatting and Pre-processing);
2. Specify the model. If you already have estimations of the pure component concentrations or spectra, enter them as ***Initial guess***. Remember to define relevant constraints: ***non-negative*** concentrations is usual, the spectra are also often ***non-negative***, while ***unimodality*** and ***closure*** may or may not apply to your case. Finally, you may also tune the ***sensitivity to pure components*** before launching the calculations;
3. View the results and choose the ***number of components*** to interpret, according to the plots of Total residuals;
4. ***Diagnose*** the model, using Sample residuals and Variable residuals;
5. ***Interpret*** the plots of Estimated Concentrations and Estimated Spectra.

Run An MCR

When your data table is displayed in the ***Editor***, you may access the **Task** menu to run a suitable analysis – for instance, MCR.

- **Task - MCR**: Run a Multivariate Curve Resolution on the current data table

Save And Retrieve MCR Results

Once the MCR has been computed according to your specifications, you may either View the results right away, or Close (and Save) your MCR result file to be opened later in the ***Viewer***.

Save Result File from the Viewer

- **File - Save**: Save result file for the first time, or with existing name
- **File - Save As**: Save result file under a new name

Open Result File into a new Viewer

- **File - Open**: Open any file or just lookup file information
- **Results - MCR**: Open MCR result file or just lookup file information
- **Results - All**: Open any result file or just lookup file information, warnings and variances

View MCR Results

Display MCR results as plots from the ***Viewer***. Your MCR results file should be opened in the Viewer; you may then access the **Plot** menu to select the various results you want to plot and interpret.

From the **View**, **Edit** and **Window** menus you may use more options to enhance your plots and ease result interpretation.

How To Plot MCR Results

- **Plot - MCR Overview**: Display the 4 main MCR plots
- **Plot - Estimated Concentrations**: Plot estimated concentrations of the chosen pure components for all samples
- **Plot - Estimated Spectra**: Plot estimated spectra of the chosen pure components
- **Plot - Residuals**: Display various types of residual plots. There you may choose between
  . MCR Fitting: Plot Sample residuals, Variable Residuals or Total residuals in your MCR model, for a

selected number of components
. PCA Fitting: Plot Sample residuals, Variable Residuals or Total residuals in a PCA model of the same data

PC Navigation Tool

Navigate up or down the PCs in your model along the vertical and horizontal axes of your plots:

- **View - Source - Back to Suggested PC**
- **View - Source - Previous Horizontal PC**
- **View - Source - Next Horizontal PC**

More Plotting Options

- **View - Source**: Select which sample types / variable types / variance type to display
- **Edit - Options**: Format your plot
- **Edit - Insert Draw Item**: Draw a line or add text to your plot
- **View - MCR Message List**: Display list of recommendations issued during the analysis, to help you improve your MCR model
- **View - Toolbars**: Select which groups of tools to display on the toolbar
- **Window - Identification**: Display curve information for the current plot

How To Change Plot Ranges:

- **View - Scaling**
- **View - Zoom In**
- **View - Zoom Out**

How To Keep Track of Interesting Objects

- **Edit - Mark**: Several options for marking samples or variables

How To Display Raw Data

- **View - Raw Data**: Display the source data for the analysis in a slave Editor

Run New Analyses From The Viewer

In the ***Viewer***, you may not only **Plot** your MCR results; the **Edit - Mark** menu allows you to mark samples or variables that you want to keep track of (they will then appear marked on all plots), while the **Task - Recalculate…** options make it possible to re-specify your analysis without leaving the viewer.

Check that the currently active subview contains the right type of plot (samples or variables) before using **Edit - Mark**.

How To Keep Track of Interesting Objects

- **Edit - Mark - One By One**: Mark samples or variables individually on current plot

- **Edit - Mark - With Rectangle**: Mark samples or variables by enclosing them in a rectangular frame (on current plot)

How To Remove Marking

- **Edit - Mark - Unmark All**: Remove marking for all objects of the type displayed on current plot

How To Reverse Marking

- **Edit - Mark - Reverse Marking**: Exchange marked and unmarked objects on the plot

How To Re-specify your Analysis

- **Task - Recalculate with Marked**: Recalculate model with only the marked samples / variables
- **Task - Recalculate without Marked**: Recalculate model without the marked samples / variables

Extract Data From The Viewer

From the ***Viewer***, use the **Edit - Mark** menu to mark samples or variables that you have reason to single out, e.g “dominant variables” or “outlying samples”, etc.

There are two ways to display the source data for the currently viewed analysis into a new Editor window.

1. Command **View - Raw Data** displays the source data into a ***slave*** Editor table, which means that marked objects on the plots result in highlighted rows (for marked samples) or columns (variables) in the Editor. If you change the marking, the highlighting will be updated; if you highlight different rows or columns, you will see them marked on the plots.
2. You may also take advantage of the **Task - Extract Data…** options to display raw data for only the samples and variables you are interested in. A new data table is created and displayed in an independent Editor window. You may then edit or re-format those data as you wish.

How To Mark Objects

Lookup the previous section Run New Analyses From The Viewer.

How To Display Raw Data

- **View - Raw Data**: Display the source data for the analysis in a slave Editor

How To Extract Raw Data

- **Task - Extract Data from Marked**: Extract data for only the marked samples / variables
- **Task - Extract Data from Unmarked**: Extract data for only the unmarked samples / variables

Three-way Data Analysis

Principles of Three-way Data Analysis

By Prof. Rasmus Bro, Royal Veterinary and Agricultural University (KVL), Copenhagen, Denmark.

If you have three-way data that is not easily described with a "flat" table structure, read about the exciting method to analyze those data (NPLS) using three-way data analysis. Before describing this tool though, it is instructive to learn what three-way data actually is and how it arises.

From Matrices and Tables to Three-way Data

In multivariate data analysis, the common situation is to have a *table* of data which is then mathematically stored in a *matrix*. All the preceding chapters have dealt with such data and in fact the whole point of linear algebra is to provide a mathematical language for dealing with such tables of data.

In some situations it is difficult to organize the data logically in a data table and the need for more complex data structures is apparent. Alongside with more complicated data it is a natural desire to be able to analyze such structures in a straightforward manner. Three-way data analysis provides one such option.

Suppose that the (e.g. spectral) measurements of a specific sample read at seven variables are given as shown below:

0.17	0.64	1.00	0.64	0.17	0.02	0.00

Thus, the data from one sample can be held in a vector. Data from several samples can then be collected in a matrix and analyzed for example with PCA or PLS.

Suppose instead that this spectrum is measured not once, but several times under different conditions. In this situation, the data may read:

0.02	0.06	0.10	0.06	0.02	0.00	0.00
0.08	0.32	0.50	0.32	0.08	0.01	0.00
0.17	0.64	1.00	0.64	0.17	0.02	0.00
0.05	0.19	0.30	0.19	0.05	0.01	0.00
0.03	0.13	0.20	0.13	0.03	0.00	0.00

where the third row is seen to be the same as above. In this case, every sample yields a table in itself. This is shown graphically as follows:

Typical sample in two-way and three-way analyses

Typical sample in two-way analysis

0.17	0.64	1.00	0.64	0.17	0.02	0.00

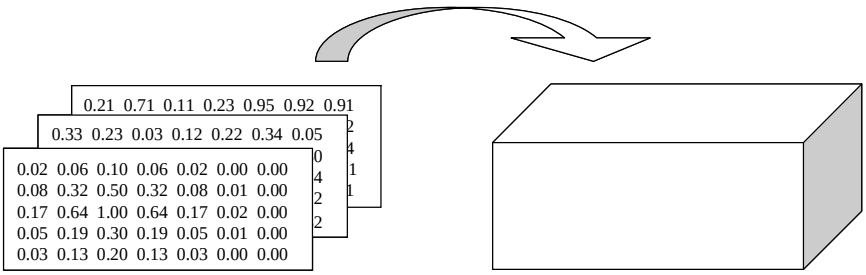
Typical sample in three-way analysis

0.02	0.06	0.10	0.06	0.02	0.00	0.00
0.08	0.32	0.50	0.32	0.08	0.01	0.00
0.17	0.64	1.00	0.64	0.17	0.02	0.00
0.05	0.19	0.30	0.19	0.05	0.01	0.00
0.03	0.13	0.20	0.13	0.03	0.00	0.00

When the data from one sample can be held in a vector, it is sometimes referred to as first-order data as opposed to scalar data – one measurement per sample – which is called zeroth-order data. When data of one sample is a matrix, then the data is called second-order data (see the 1988 article by Sanchez and Kowalski – detailed bibliography given in the Method References chapter).

Having several sets of matrices, for example from different samples, a ***three-way array*** is obtained (see figure below). Three-way data analysis is the analysis of such structures.

A three-way array is obtained from several sets of matrices



In the same way as going from two-way matrices to three-way arrays, it is also possible to obtain four-way, five-way, or multi-way in general, data. Multi-way data is sometimes referred to as N-way data, which is where the N in NPLS (see below) comes from.

Notation of Three-way Data

In order to be able to discuss the properties of three-way data and the models built from them, a proper notation is needed. A suggestion for general multi-way notation has been offered in the literature, see for instance Kiers 2000 (detailed bibliography given in the Method References chapter). Some minor modifications and additions will be made here, but all in all, it is useful to use the suggested notation as it will also make it easier to absorb the general literature on multi-way analysis.

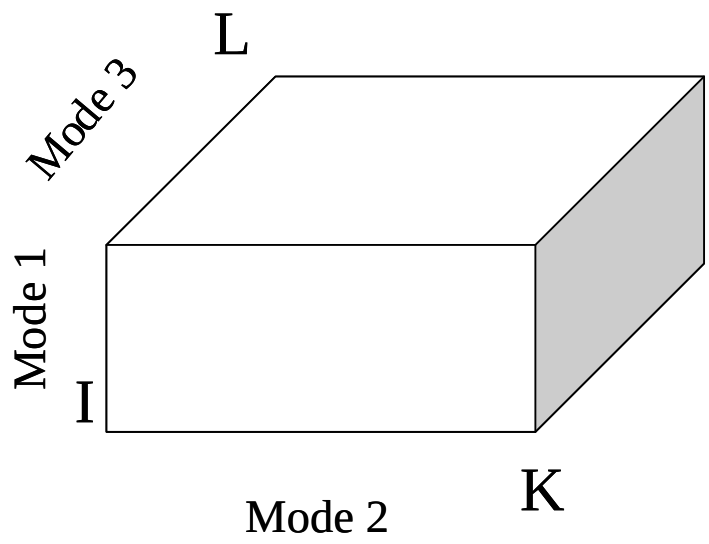
Modes of a Three-way Array

A three-way array can also be called a third-order tensor or a multimode array, but the former is preferred here. Sometimes in psychometric literature a distinction is made between **modes** and ways, but this is not needed

here. Note that a three-way array is not referred to as a three-dimensional array. The term dimension is retained for indicating the size of each mode.

The definition of which is the first, second and third mode can be seen in the figure below. The dimensions of these modes are I, K and L respectively.

First, second and third modes in a three-way array



Two different types of modes will be distinguished. One is a sample-mode and the other is a variable-mode.

For a typical two-way (matrix) data set, the samples are held in the first (row) mode and the variables are held in the second (column) mode. This configuration is also sometimes called OV where O means that the first mode is an object-mode and V means that the second mode is a variable mode. If a grey-level image is analyzed and the image represents a measurement on a sample, then the matrix holding the data is a V^2 structure because both modes represent different measurements on the same sample.

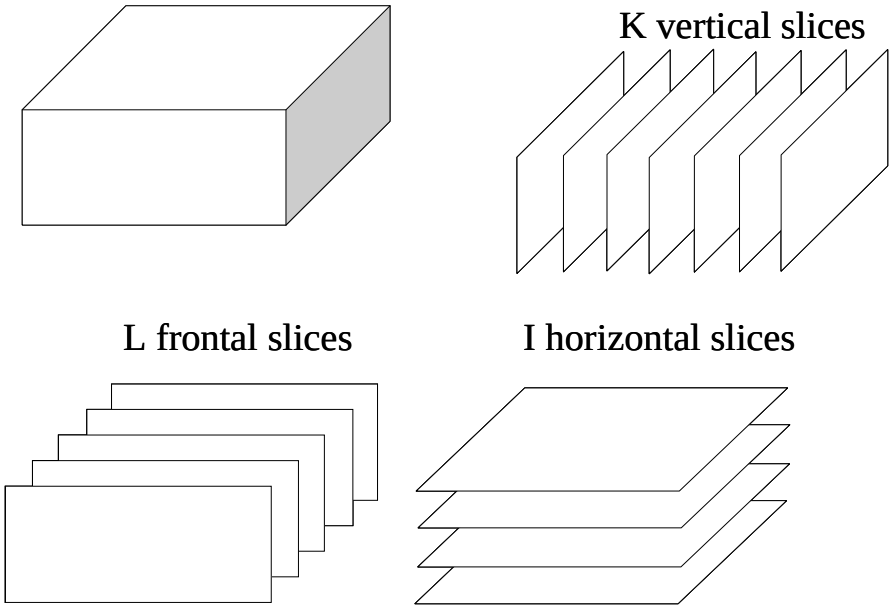
Likewise, for three-way data, several types of structures such as OV^2, O^2V, V^3 etc. can be imagined. In the following, only OV^2 data are considered in detail.

Note: As in two-way analysis it is common practice to keep samples in the first mode for OV^2 data.

Substructures in Three-way Arrays

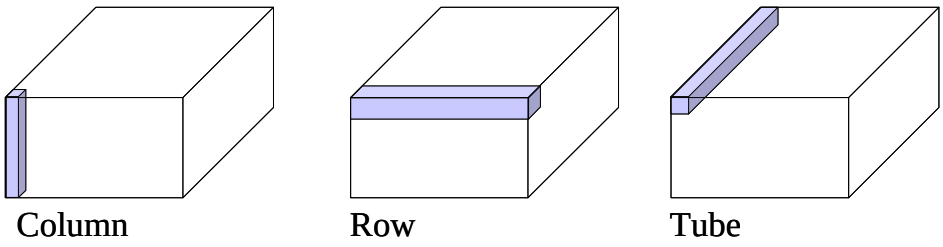
A two-way array can be divided into individual columns or into individual rows. A three-way array can be divided into frontal, horizontal or vertical slices (matrices):

Frontal, horizontal and vertical slices of a three-way array



It is also possible to divide further into vectors. Rather than just rows and columns, there are rows, columns and tubes as shown below.

Rows, columns and tubes in a three-way array



Types of Three-way Data

So where do three-way data occur? As a matter of fact, it occurs more often than one may anticipate. Some examples will illustrate this.

Examples:

- *Infrared spectra* (300 wavelengths) are measured on several samples (50). A spectrum is measured on each sample at five distinct temperatures. In this case, the data can be arranged as a 50×300×5 array.
- The concentrations of seven chemical species are determined weekly at 23 locations in a lake for one year in an *environmental analysis*. The resulting data is a 23×7×52 array.
- In a *sensory experiment*, eight assessors score on 18 different attributes on ten different sorts of apples. The data can consequently be arranged in 10×8×18 array.
- Seventy-two samples are measured using *fluorescence* excitation-emission spectroscopy with 100 excitation wavelengths and 540 emission wavelengths. The excitation-emission data can be held in 72×540×100 array.
- Twelve batches are monitored with respect to nine *process* variables every minute for two hours. The data are arranged as a 12×9×120 array.

- Fifteen food samples have been assessed using texture-measurements (40 variables) after six different types of storage conditions. The subsequent data can be stored in a 15×40×6 array.

As can be seen, many types of data are conveniently seen as three-way data.

Note: There is no practical consequence of whether the second and third modes are interchanged. As long as samples are kept in the first mode, the choice between the second and third mode is immaterial except for the trivial interchanged interpretation.

Is a Three-way Structure Appropriate for my Data?

It is worth also to consider what are not appropriate three-way data sets. A simple example: A two-way data set is obtained of size 15 (samples) × 50 (variables). Now this matrix is duplicated yielding another identical matrix. Even though this combined data set can be arranged as a three-way 15×40×2 array, it is evident that no new information is obtained by doing so. So, although the data are three-way data, no added value is expected by this modified representation. What then if the additional data set was not a duplicate but a replicate, hence a re-measured data set? Then indeed, the two matrices are different and can more meaningfully be arranged as a three-way data set. But imagine a set of samples where one variable is measured several times. Even though the replicate measurements can be arranged in a two-way matrix and analyzed e.g. with PCA, it will usually not yield the most interesting results as all the variables are hopefully identical up to noise. In most cases, such data are better analyzed by seeing the replicates as new samples. Then the score plots will reveal any differences between individual measurements. Likewise, a set of replicate matrices are mostly better analyzed with two-way methods.

Another important example on something that is not feasible with three-way data is the following. If a set of NIR spectra (100 variables) is measured alongside with Ultraviolet Visible (UVVis) spectra (100 variables), then it is not feasible to join the two matrices in a three-way array. Even though the sizes of the two matrices fit together, there is no correspondence between the variables and hence such a three-way array makes no sense. Such data are two-way data: the two matrices have to be put next to each other, just like any other set of variables are held in a matrix.

Three-way Regression

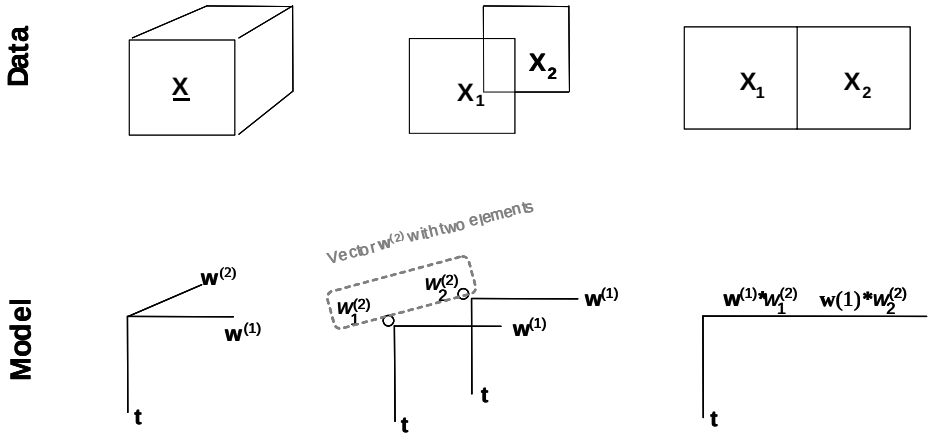
With a three-way array $\underline{\mathbf{X}}$ and matrix $\mathbf{Y}$ or vector $\mathbf{y}$ it is possible to build three-way regression models. The principle in three-way regression is more or less the same as in two-way regression. The regression method N-PLS is the extension of ordinary PLS to arbitrary ordered data. For three-way data specifically, the term tri-PLS is used. Tri-PLS provides a model of $\mathbf{X}$ which predicts the dependent variable $\mathbf{Y}$ through an inner relation just like in two-way PLS.

The model of $\underline{\mathbf{X}}$ is a trilinear model which is easily shown graphically, but complicated to write in matrix notation. Matrices are intrinsically connected to two-way data, so in order to write a three-way model in matrices, the data and the model have to be rearranged into a two-way model. For appropriately pre-processed data (See chapter Pre-processing of Three-way data) the tri-PLS model consists of a model of $\underline{\mathbf{X}}$, a model of $\mathbf{Y}$ and an inner relation connecting these.

One-component Tri-PLS Model of X-data

The figure below shows how a three-way data set and associated trilinear model can be represented as matrices. The three-way data set $\underline{\mathbf{X}}$ has only two frontal slices in this case, i.e. dimension two in the third mode for simplicity. By putting these two frontal slices next to each other a two-way matrix is obtained. This representation of the data does not change the actual content of the array but merely serves to enable standard linear algebra to be used here. The data can now be written as a two-way (dim $I*KL$) matrix $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$.

Principle in rearranging a three-way array and the corresponding one-component trilinear model to matrix-form.



A one-component model of $\underline{\mathbf{X}}$ is also shown. More components are easily added, but one is enough to show the principle of the rearranging. The trilinear component consists of a score vector $\mathbf{t}$ (dim I*1), a weight vector in the first variable mode $\mathbf{w}^{(1)}$ (dim K*1) and a weight vector in the second variable mode $\mathbf{w}^{(2)}$ (dim L*1). These three vectors can be rearranged similarly to the data leading to a matrix representation of the trilinear component which can then be written

$$\hat{\mathbf{X}} = \mathbf{t}\begin{bmatrix} \mathbf{w}^{(1)} * w_1^{(2)} \\ \mathbf{w}^{(1)} * w_1^{(2)} \end{bmatrix}^T = \mathbf{t}(\mathbf{w}^{(2)} \otimes \mathbf{w}^{(1)})^T$$

where the Kronecker product is used to abbreviate the expression in parentheses. While this two-way representation looks a bit complicated, it is noteworthy that it simply expresses the trilinear model shown in the above figure using two-way notation. Additionally, it represents the trilinear model as a bilinear model using a score vector and a vector combined from the two weight vectors.

Only Weights and no Loadings

In tri-PLS there are no loadings introduced. In essence, loadings are introduced in two-way PLS to provide orthogonal scores. However, the introduction of multi-way loadings will not give orthogonal scores and these loadings are therefore not needed (see Bro 1996 and Bro & al. 2001 - detailed bibliography given in the Method References chapter, which is available as a .PDF file from CAMO's web site www.camo.com/TheUnscrambler/Appendices).

An A-component Tri-PLS Model of X-data

When there is more than one component in the tri-PLS model of the data, a so-called core array is added. This core array is a computational construct which is found after the whole model has been fitted. It does not affect the predictions at all but only serves to provide an adequate model of $\underline{\mathbf{X}}$ hence adequate residuals. The purpose of this core is to take possible interactions between components into account. Because the scores and weight vectors are not orthogonal (See Section Non-orthogonal Scores and Weights), it is possible that a better fit to $\underline{\mathbf{X}}$ can be obtained by allowing for example score one to interact with weight two etc. This introduction of interactions is usually not considered when validating the model. It is simply a way of obtaining more reasonable X-residuals (see Bro & al. 2001 - detailed bibliography given in the Method References chapter). When the model has been found, only scores, weights and residuals are used for investigating the model as is the case in two-way PLS.

The A-component tri-PLS model of $\underline{\mathbf{X}}$ can be written

$$\hat{\mathbf{X}} = \mathbf{TG}(\mathbf{W}^{(2)} \otimes \mathbf{W}^{(1)})^T$$

where the rearranged matrix **G** is originally the (dim *A**A**A*) core array that takes possible interactions into account.

The Inner Relation

Just like in two-way PLS, the *inner relation* is the core of tri-PLS model. Scores in **X** are used to predict the scores in **Y** and from these predictions, the estimated $\hat{\mathbf{Y}}$ is found. This connection between **X** and **Y** through their scores is called the inner relation. It consists of a regression step, where the scores in **X** are used for predicting the scores in **Y**. Thus, from a new sample we can predict its correspond **Y** scores. As a model of **Y** is given by the scores times the loadings, we can predict the unknown **Y** from these estimated scores.

Because the scores are not orthogonal in tri-PLS, the inner relation is a bit different from the ordinary two-way case. When predicting the *a*'th score of **Y**, *all* scores from 1 to *a* in **X** have to be taken into account. Therefore,

$$\hat{\mathbf{u}}_a = \mathbf{T}_{1\text{-}a}\hat{\mathbf{b}}_{1\text{-}a}$$

where $\mathbf{T}_{1\text{-}a}$ is a matrix containing all the first *a* score vectors.

The Prediction Step

The prediction of **Y** is simply found from the predicted scores and the prior Y-loadings as

$$\hat{\mathbf{Y}} = \hat{\mathbf{U}}\mathbf{Q}^T .$$

Main Results of Tri-PLS Regression

The interpretation of a tri-PLS model is similar to a two-way PLS model because most of the results are expressed in a similar way. There are scores, weights, regression coefficients and residuals. All of these are interpreted in much the same way as in ordinary PLS (see Chapter Main Results of Regression p. 111 for more details). Only the main differences are highlighted in the following.

No Loadings in tri-PLS

As mentioned in chapter Three-way Regression (see for instance section Only Weights and no Loadings), a tri-PLS model is expressed with two sets of weights (similar to the loading weights in PLS) but no loadings are computed. Thus the interpretation of tri-PLS results will, as far as the Predictor variables are concerned, focus on the X-weights.

Two Sets of X-weights in tri-PLS

In tri-PLS there are weights for the first *and* the second variable mode. Assume, as an example, that a data set is given with wavelengths in variable mode one and with different times in variable mode two.

If the weights in variable mode one are high for, for example, the first and third wavelengths, then, as in two-way PLS, these wavelengths influence the model more than the others. Unlike two-way PLS, the weights in one mode, however, do not provide the whole story. Even though wavelength one and three in variable mode one are high, their total impact on the model has to be viewed based on the weights in variable mode two.

If only one specific time has high weights in variable mode two, then the high impact of wavelength one and three is primarily due to the variation at that specific time in variable mode two. Therefore, if that particular time is actually representing an erroneous set of measurements, then the relative influences in the wavelength mode may change completely upon deletion of that time in variable mode two.

Non-orthogonal Scores and Weights

Orthogonality properties of scores and weights are seldom of too much practical concern in PLS regression. Orthogonality is primarily important in the mathematical derivations and in developing algorithms.

In some situations, the non-orthogonal nature of scores and weights in tri-PLS may lead to surprising, though correct, models. For example, two weight vectors of two different components may turn out very similar. This can happen if the same variation in one variable mode is related to two different phenomena in the data.

For instance, a general increase over time (variable mode one) may occur for two different spectrally detected substances (variable mode two). In such a case, the appearance of two similar weight vectors is merely a useful flagging of the fact that the same time-trend affects different parts of the model.

Maximum Number of Components

The formula for determining the maximum possible number of components in PLS1 and PLS2 is min (I-1, K) with I the number of samples in the calibration set and K the number of variables. In Three-way PLS there are two variable modes, such that the maximum possible number of components is min(I-1, K*L) with K and L the numbers of primary and secondary variables. If the data is not centered, the maximum number of components is min(I,K*L).

Interpretation of a Tri-PLS Model

Once a three-way regression model is built, you have to ***diagnose*** it, i.e. assess its quality, before you can start ***interpreting*** the relationship between X and Y. Finally, your model will be ready for use for ***prediction*** once you have thoroughly checked and refined it.

Most tri-PLS results are interpreted in much the same way as in ordinary PLS (see Chapter “Main Results of Regression” p. 111 for more details). Exceptions are listed in Chapter “Main Results of Tri-PLS Regression” above.

Read more about specific details:

- Interpretation of variances p. 101
- Interpretation of the two sets of weights p. 183
- Interpretation of non-orthogonal scores and weights p. 184
- How to detect outliers in regression p. 115

Three-way Data Analysis in Practice

The sections that follow list menu options, dialogs and plots for three-way data analysis (nPLS). For a more detailed description of each menu option, read The Unscrambler Program Operation, available as a PDF file from Camo’s web site www.camo.com/TheUnscrambler/Appendices.

In practice, building and using a tri-PLS regression model consists of several steps:

1. Choose and implement an appropriate ***pre-processing*** method. Individual modes of a 3-D data array may be transformed in the same way as a “normal” data vector (see Chapter Re-formatting and Pre-processing);
2. Build the model: ***calibration*** fits the model to the available data, while ***validation*** checks the model for new data;
3. Choose the ***number of components*** to interpret, according to calibration and validation variances;

4. ***Diagnose*** the model, using variance curves, X-Y relation outliers, Predicted vs. Measured;
5. ***Interpret*** the scores and weights plots and the B-coefficients;
6. ***Predict*** response values for new data (optional).

Run A Tri-PLS Regression

When your 3-D data table is displayed in the ***Editor***, you may access the **Task** menu to run a suitable analysis – here, tri-PLS Regression.

- **Task - Regression**: Run a tri-PLS regression on the current 3-D data table

Save And Retrieve Tri-PLS Regression Results

Once the tri-PLS regression model has been computed according to your specifications, you may either View the results right away, or Close (and Save) your results as a Three Way PLS file to be opened later in the ***Viewer***.

Save Result File from the Viewer

- **File - Save**: Save result file for the first time, or with existing name
- **File - Save As**: Save result file under a new name

Open Result File into a new Viewer

- **File - Open**: Open any file or just lookup file information
- **Results - Regression**: Open regression result file or just lookup file information, warnings and variances
- **Results - All**: Open any result file or just lookup file information, warnings and variances

View Tri-PLS Regression Results

Display Three Way PLS results as plots from the ***Viewer***. Your Three Way PLS results file should be opened in the Viewer; you may then access the **Plot** menu to select the various results you want to plot and interpret.

From the **View**, **Edit** and **Window** menus you may use more options to enhance your plots and ease result interpretation.

How To Plot tri-PLS Regression Results

- **Plot - Regression Overview**: Display the 4 main regression plots
- **Plot - Variances and RMSEP**: Plot variance curves
- **Plot - Sample Outliers**: Display 4 plots for diagnosing outliers
- **Plot - X-Y Relation Outliers**: Display t vs. u scores along individual PCs
- **Plot - Scores and Loading Weights**: Display scores and weights separately or as a bi-plot
- **Plot - Predicted vs Measured**: Display plot of predicted Y values against actual Y values
- **Plot - Scores**: Plot scores along selected PCs
- **Plot - Loading Weights**: Plot loading weights along selected PCs
- **Plot - Important Variables**: Display 2 plots to detect most important variables

- **Plot - Regression Coefficients**: Plot regression coefficients
- **Plot - Regression and Prediction**: Display Predicted vs. Measured and Regression coefficients
- **Plot - Residuals**: Display various types of residual plots
- **Plot - Leverage**: Plot sample leverages

For more options allowing you to re-format your plots, navigate along PCs, mark objects etc., look up chapter View PCA Results p. 103 All the menu options shown there also apply to regression results.

Run New Analyses From The Viewer

In the ***Viewer***, you may not only **Plot** your Three Way PLS results; the **Edit - Mark** menu allows you to mark samples or variables that you want to keep track of (they will then appear marked on all plots), while the **Task - Recalculate…** options make it possible to re-specify your analysis without leaving the viewer.

Check that the currently active subview contains the right type of plot (samples or variables) before using **Edit - Mark**.

Look up the relevant menu options in chapter Run New Analyses from the Viewer (for PCA) p. 104.Most of the menu options shown there also apply to three-way regression results.

Extract Data From The Viewer

From the ***Viewer***, use the **Edit - Mark** menu to mark samples or variables that you have reason to single out, e.g "significant X-variables" or "outlying samples", etc.

Look up details and relevant menu options in chapter Extract Data from the Viewer (for PCA) p. 105. Most of the menu options shown there also apply to regression results.

How to Run Other Analyses on 3-D Data

The only option in the **Task** menu available for 3-D data is **Task - Regression**. Other types of analysis apply to 2-D data only.

Useful tips

To run an analysis (other than three-way regression) on your 3-way data, you need to ***duplicate your 3-D table as 2-D data*** first. Then all relevant analyses will be enabled.

For instance, you may run an exploratory analysis with PCA on unfolded 3-way spectral data, by doing the following sequence of operations:

1. Start from your 3-D data table (OV^2 layout) where each row contains a 2-way spectrum;
2. Use **File - Duplicate - As 2-D Data Table**: this generates a 2-D table containing unfolded spectra;
3. Save the resulting 2-D table with **File - Save As**;
4. Use **Task - PCA** to run the desired analysis.

Another possibility is to develop your own three-way analysis routine and implement it as a User-Defined Analysis (UDA). Such analyses may then be run from the **Task - User-defined Analysis** menu .

Interpretation Of Plots

This chapter presents all predefined plots available in The Unscambler. They are sorted by plot types:

- Line;
- 2D Scatter;
- 3D Scatter;
- Matrix;
- Normal Probability;
- Table plots;
- Special plots.

Whenever viewing a plot in The Unscrambler, hitting <F1> will display the Help chapter on how to interpret the type of plot which is currently active in your viewer.

Line Plots

Detailed Effects (Line Plot)

This plot displays all effects for a given response variable. It is recommended to choose a layout as **bars** to make it easier to read. Each effect (main effect, interaction) is represented by a bar.

A bar pointing upwards indicates a positive effect. A bar pointing downwards indicates a negative effect. Click on a bar to read the exact value of the calculated effect.

Discrimination Power (Line Plot)

This plot shows how much each X-variable contributes to separating two classes. There must always be some variables with good discrimination power in order to achieve good classifications.

A discrimination power near 1 indicates that the variable concerned is of no use when it comes to separating the two classes. A discrimination power larger than three indicates an important variable.

Variables with low discrimination power and low modeling power do not contribute to the classification: you should go back to your class models and refine them by keeping out those variables.

Estimated Concentrations (Line Plot)

This plot, available for MCR results, displays the estimated concentrations of two or more constituents across all the samples included in the analysis.

Each plotted curve is the estimated concentration profile of one given constituent.

The curves are plotted for a fixed number of components in the model; note that in MCR, the number of model dimensions (components) also determines the number of resolved constituents. Therefore, if you tune the

number of PCs up or down with the toolbar buttons or , this will also affect the number of curves displayed.

For instance, if the plot currently displays 2 curves, clicking will update the plot to 3 curves representing the profiles of 3 constituents in a 3-dimensional MCR model.

Estimated Spectra (Line Plot)

This plot, available for MCR results, displays the estimated spectra of two or more constituents across all the variables included in the analysis.

Each plotted curve is the estimated spectrum of one pure constituent.

The curves are plotted for a fixed number of components in the model; note that in MCR, the number of model dimensions (components) also determines the number of resolved constituents. Therefore, if you tune the number of PCs up or down with the toolbar buttons or , this will also affect the number of curves displayed.

For instance, if the plot currently displays 2 curves, clicking will update the plot to 3 curves representing the spectra of 3 constituents in a 3-dimensional MCR model.

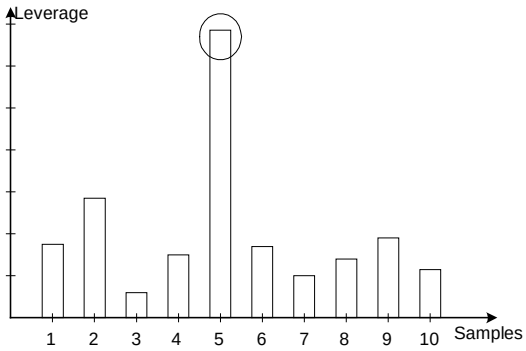
F-Ratios of the Detailed Effects (Line Plot)

This is a plot of the f-ratios of the effects in the model. F-ratios are not immediately interpretable, since their significance depends on the number of degrees of freedom. However, they can be used as a visual diagnostic: effects with high f-ratios are more likely to be significant than effects with small f-ratios.

Leverages (Line Plot)

Leverages are useful for detecting samples which are far from the center within the space described by the model. Samples with high leverage differ from the average samples; in other words, they are likely outliers. A large leverage also indicates a high influence on the model. The figure below shows a situation where sample 5 is obviously very different from the rest and may disturb the model.

One sample has a high leverage



Leverages can be interpreted in two ways: absolute, and relative.

The absolute leverage values are always larger than zero, and can go (in theory) up to 1. As a rule of thumb, samples with a leverage above 0.4 - 0.5 start being bothering.

Influence on the model is best measured in terms of relative leverage. For instance, if all samples have leverages between 0.02 and 0.1, except for one which has a leverage of 0.3, although this value is not extremely large, the sample is likely to be influential.

Leverages in Designed Data

For designed samples, the leverages should be interpreted differently whether you are running a regression (with the design variables as X-variables) or just describing your responses with PCA.

By construction, the leverage of each sample in the design is known, and these leverages are optimal, i.e. all design samples have the same contribution to the model. So do not bother about the leverages if you are running a regression: the design has cared for it.

However, if you are running a PCA on your response variables, the leverage of each sample is now determined with respect to the response values. Thus some samples may have high leverages, either in an absolute or a relative sense. Such samples are either outliers, or just samples with extreme values for some of the responses.

What Should You Do with a High-Leverage Sample?

The first thing to do is to understand why the sample has a high leverage. Investigate by looking at your raw data and checking them against your original recordings.

Once you have found an explanation, you are usually in one of the following cases.

Case 1: there is an error in the data. Correct it, or if you cannot find the true value or re-do the experiment which would give you a more valid value, you may replace the erroneous value with "missing".

Case 2: there is no error, but the sample is different from the others. For instance, it has extreme values for several of your variables. Check whether this sample is "of interest" (e.g. it has the properties you want to achieve, to a higher degree than the other samples), or "not relevant" (e.g. it belongs to another population than the one you want to study). In the former case, you will have to try to generate more samples of the same kind: they are the most interesting ones! In the latter case (and only then), you may remove the high-leverage sample from your model.

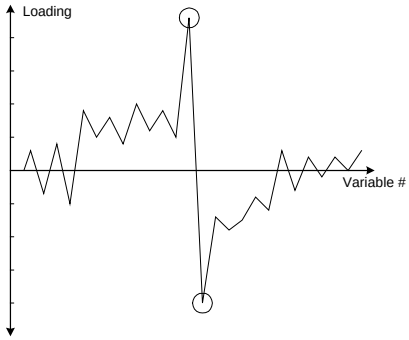
Loadings for the X-variables (Line Plot)

This is a plot of X-loadings for a specified component versus variable number. It is useful for detecting important variables. In many cases it is usually better to look at two- or three-vector loading plots instead because they contain more information.

Line plots are most useful for multi-channel measurements, for instance spectra from a spectrophotometer, or in any case where the variables are implicit functions of an underlying parameter, like wavelength, time,…

The plot shows the relationship between the specified component and the different X-variables. If a variable has a large positive or negative loading, this means that the variable is important for the component concerned; see the figure below**.** For example, a sample with a large score value for this component will have a large positive value for a variable with large positive loading.

Line plot of the X-loadings, two important variables



Variables with large loadings in early components are the ones that vary most. This means that these variables are responsible for the greatest differences between the samples.

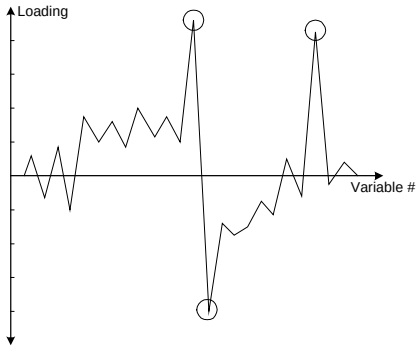
Note: Passified variables are displayed in a different color so as to be easily identified.

Loadings for the Y-variables (Line Plot)

This is a plot of Y-loadings for a specified component versus variable number. It is usually better to look at 2D or 3D loading plots instead because they contain more information. However, if you have reason to study the X-loadings as line plots, then you should also display the Y-loadings as line plots in order to make interpretation easier.

The plot shows the relationship between the specified component and the different Y-variables. If a variable has a high positive or negative loading, as in the example plot shown below, this means that the variable is well explained by the component. A sample with a large score for the specified component will have a high value for all variables with large positive loadings.

Line plot of the Y-loadings, three important variables



Y-variables with large loadings in early components are the ones that are most easily modeled as a function of the X-variables.

Note: Passified variables are displayed in a different color so as to be easily identified.

Loading Weights (Line Plot)

This is a two dimensional scatter plot of X-loading weights for two specified components from a PLS analysis. It can be useful for detecting which X-variables are most important for predicting Y, although it is better to use the 2D scatter plot of X-loading weights and Y-loadings.

Note 1: The X-loading weights for PC1 are exactly the same as the regression coefficients for PC1.

Note 2: Passified variables are displayed in a different color so as to be easily identified.

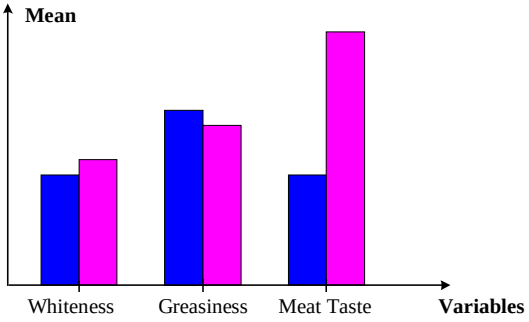
Mean (Line Plot)

For each variable, the average over all samples in the chosen sample set is displayed as a vertical bar.

If you have chosen to display groups or subgroups of samples, the plot has one bar per group (or subgroup), for each variable. You can easily compare the averages between groups.

For instance, if the data are results from designed experiments, a plot showing the average for the whole design and the average over the center samples is very useful to detect a possible curvature in the relationship between the response and the design variables. The figure below shows such an example: Responses 1 and 2 seem to have a linear relationship with the design variables, whereas for response 3 the center samples have a much higher average than the cube samples, which indicates a non-linear relationship between response 3 and some of the design variables. If this is the case at a screening stage, you should investigate further with an optimization design, in order to fit a quadratic response surface.

Mean for 3 responses, with groups "Design samples" and "Center samples"



Model Distance (Line Plot)

This plot visualizes the distance between one class and all other classes (models) used in the classification.

The distance from a class (model) to itself is by definition 1.0. The distance to other classes should be greater than three for good separation between classes.

Modeling Power (Line Plot)

The Modeling Power plot is used to study the relevance of a variable. It tells you how much of the variable's variance is used to describe the class (model).

Modeling power is always between 0 and 1. A variable with a modeling power higher than 0.3 is important in modeling what is typical of that class.

Variables with low discrimination power and low modeling power do not contribute to the classification: you should go back to your class models and refine them by keeping out those variables.

Predicted and Measured (Line Plot)

In this plot, you find the measured and predicted Y-values plotted in parallel for each sample. You can spot which samples are well predicted and which ones are not. If necessary, try transforming your data table or removing outliers to make a better model. Using more components during prediction may improve the predictions, but do this only if the validated residual variance does not increase.

You should use the optimal number of components determined by validation.

p-values of the Detailed Effects (Line Plot)

This is a plot of the p-values of the effects in the model. Small values (for instance less than 0.05 or 0.01) indicate that the effect is significantly different from zero, i.e. that there is little chance that the observed effect is due to mere random variation.

p-values of the Regression Coefficients (Line Plot)

This is a plot of the p-values for the different regression coefficients (B). Small values (for instance less than 0.05 or 0.01) indicate that the corresponding variable has a significant effect on the response (given that all the other variables are present in the model).

Regression Coefficients (Line Plot)

Regression coefficients summarize the relationship between all predictors and a given response. For PCR and PLS, the regression coefficients can be computed for any number of components. The regression coefficients for 5 PCs, for example, summarize the relationship between the predictors and the response, as it is approximated by a model with 5 components.

Note: What follows applies to a line plot of regression coefficients in general. To read about specific features related to three-way PLS results, look up the Details section below.

This plot shows the regression coefficients for one particular response variable (Y), and for a model with a particular number of components. Each predictor variable (X) defines one point of the line (or one bar of the plot). It is recommended to configure the layout of your plot as **bars**.

The regression coefficients line plot is available in two options: weighted coefficients (BW), or raw coefficients (B). The respective constant values B0W or B0 are indicated at the bottom of the plot, in the Plot ID field (use **View - Plot ID**).

Note: The weighted coefficients (BW) and raw coefficients (B) are identical if no weights where applied on your variables.

If you have weighted your predictor variables with 1/Sdev (standardization), the ***weighted*** regression coefficients (BW) take these weights into account. Since all predictors are brought back to the same scale, the coefficients show the relative importance of the X-variables in the model.

The ***raw*** coefficients are those that may be used to write the model equation in original units:

Y = B0 + B1 * X-variable1 + B2 * X-variable2 + …

Since the predictors are kept in their original scales, the coefficients do not reflect the relative importance of the X-variables in the model.

Weighted Regression Coefficients (Bw)

Predictors with a large regression coefficient play an important role in the regression model; a positive coefficient shows a positive link with the response, and a negative coefficient shows a negative link.

Predictors with a small coefficient are negligible. You can mark them and recalculate the model without those variables.

Raw Regression Coefficients (B)

The main application of the raw regression coefficients is to build the model equation in original units.

The raw coefficients do **not** reflect the importance of the X-variables in the model, because the sizes of these coefficients depend on the range of variation (and indirectly, on the original units) of the X-variables. A small raw coefficient does not necessarily indicate an unimportant variable; a large raw coefficient does not necessarily indicate an important variable.

If your purpose is to identify important predictors, always use the weighted regression coefficients plot if you have standardized the data. If not, use plots with t-values and p-values when available (for MLR and Response Surface).

Last, you may alternatively display the Uncertainty Limits (for PCR and PLS), which are available if you used Cross-Validation and the Uncertainty Test option in the **Regression** dialog.

Line Plot of Regression Coefficients: Three-Way PLS

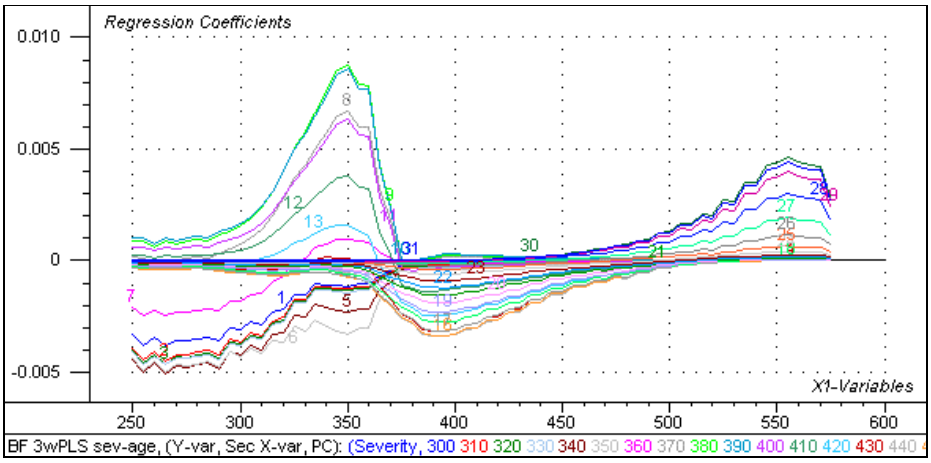
In a three-way PLS model, each Y-variable is modeled as a function of the combination of Primary and Secondary X-variables. Thus the relationship between Y and X1 can be expressed with an equation (using regression coefficients) that varies as a function of X2 – and vice-versa.

As a consequence, the line plots of regression coefficients are available in two versions:

- With all X1-variables along the abscissa; Y is fixed (as selected in the **Regression Coefficients** plot dialog), and the plot shows one curve for each X2-variable;
- With all X2-variables along the abscissa; Y is fixed (as selected in the **Regression Coefficients** plot dialog), and the plot shows one curve for each X1-variable.

The plot can be interpreted by looking for regions in X1 (resp. X2) with large positive or negative coefficients for some or all of the X2- (resp. X1-) variables. In the example below, the most interesting X1-region with respect to response “Severity” is around 350, with three additional peaks: 250-290, 390-400 and 550-560.

Line plot of X1-Regression Coefficients for response Severity



Regression Coefficients with t-values (Line Plot)

Regression coefficients (B) are primarily used to check the importance of the different X-variables in predicting Y. Large absolute values indicate large importance (significance) and small values (close to 0)

indicate an unimportant variable. The coefficient value indicates the average increase in Y when the corresponding X-variable is increased by one unit, keeping all other variables constant.

The critical value for the different regression coefficients (5% level) is indicated by a straight line. A coefficient with a larger absolute value than the straight line, is significant in the model.

The plots of the t- and p-values for the different coefficients may also be added.

RMSE (Line Plot)

This plot gives the square root of the residual variance for individual responses, back-transformed into the same units as the original response values. This is called

- **RMSEC** (Root Mean Square Error of Calibration) if you are plotting Calibration results;
- **RMSEP** (Root Mean Square Error of Prediction) if you are plotting Validation results.

The RMSE is plotted as a function of the number of components in your model. There is one curve per response (or two if you have chosen Cal and Val together). You can detect the optimal number of components: this is where the Val curve (i.e. RMSEP) reaches a minimum.

Sample Residuals, MCR Fitting (Line Plot)

This plot displays the residuals for each sample for a given number of components in an MCR model.

The size of the residuals is displayed on the scale of the vertical axis. The plot contains one point for each sample included in the analysis; the samples are listed along the horizontal axis.

The sample residuals are a measure of the distance between each sample and the MCR model. Each sample residual varies depending on the number of components in the model (displayed in parentheses after the name of the model, at the bottom of the plot). You may tune up or down the number of components for which the residuals are displayed, using the or toolbar buttons.

The size of the residuals tells you about the misfit of the model. It may be a good idea to compare the sample residuals from an MCR fitting to a PCA fit on the same data (displayed on the plot of Sample Residuals, PCA Fitting). Since PCA provides the best possible fit along a set of orthogonal components, the comparison tells you how well the MCR model is performing in terms of fit.

Note that, in the MCR Overview, both plots are displayed side by side in the lower part of the Viewer. Check the scale of the vertical axis on each plot to compare the sizes of the residuals.

Sample Residuals, PCA Fitting (Line Plot)

This plot is available when viewing the results of an MCR model. It displays the sample residuals from a PCA model on the same data.

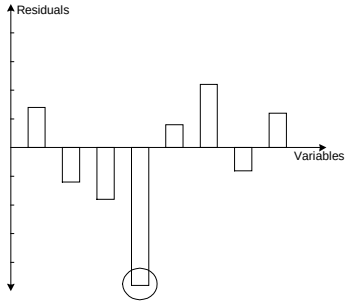
This plot is supposed to be used as a basis for comparison with the Sample Residuals, MCR fit (the actual residuals from the MCR model). Since PCA provides the best possible fit along a set of orthogonal components, the comparison tells you how well the MCR model is performing in terms of fit.

Note that, in the MCR Overview, both plots are displayed side by side in the lower part of the Viewer. Check the scale of the vertical axis on each plot to compare the sizes of the residuals.

Sample Residuals, X-variables (Line Plot)

This is a plot of the residuals for a specified sample and component number for all the X-variables. It is useful for detecting outlying sample or variable combinations. Although outliers can sometimes be modeled by incorporating more components, this should be avoided since it will reduce the prediction ability of the model.

Line plot of the sample residuals: one variable is outlying

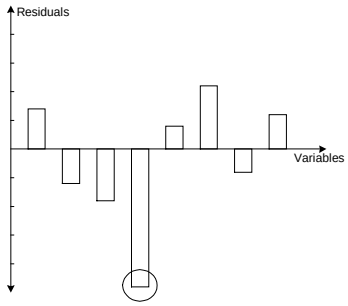


In contrast to the variable residual plot, which gives information about residuals for all samples for a particular variable, this plot gives information about all possible variables for a particular sample. It is therefore useful when studying how a specific sample fits to the model.

Sample Residuals, Y-variables (Line Plot)

A plot of the residuals for a specified sample and component number for all the Y-variables, this plot is useful for detecting outlying sample/variable combinations, as shown in the figure below. While outliers can sometimes be modeled by incorporating more components, this should be avoided since it will reduce the prediction ability of the model.

Line plot of the sample residuals: one variable is outlying



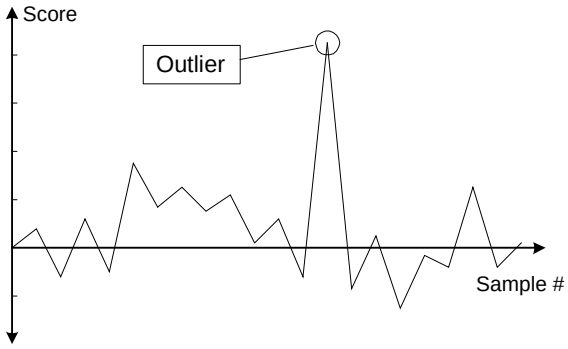
This plot gives information about all possible variables for a particular sample (as opposed to the variable residual plot, which gives information about residuals for all samples for a particular variable), and therefore indicates how well a specific sample fits to the model.

Scores (Line Plot)

This is a plot of score values versus sample number for a specified component. Although it is usually better to look at 2D or 3D score plots because they contain more information, this plot can be useful whenever the samples are sorted according to the values of an underlying variable, e.g. time, to detect trends or patterns.

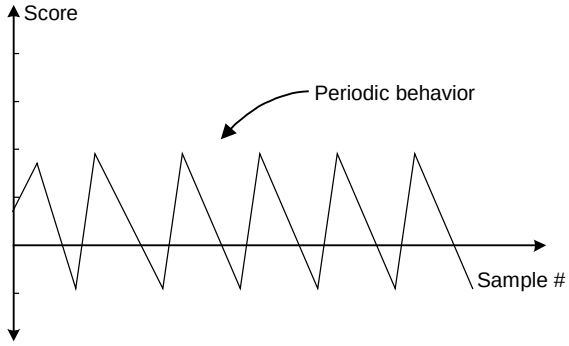
The smaller the vertical variation (i.e. the closer the score values are to each other), the more similar the samples are for this particular component. Look for samples which have a very large positive or negative score value compared to the others: these may be outliers.

An outlier sticks out on a line plot of the scores



Also look for systematic patterns, like a regular increase or decrease, periodicity, etc.… (only relevant if the sample number has a meaning, like time for instance).

Line plot of the scores for time-related data



Standard Deviation (Line Plot)

For each variable, the standard deviation (square root of the variance) over all samples in the chosen sample set is displayed.

This plot may be useful to detect which variables have the largest absolute variation. If your variables have different standard deviations, you will need to standardize them in later multivariate analyses.

Standard Error of the Regression Coefficients (Line Plot)

This is a plot of the standard errors of the different regression coefficients (B). These values can be used to compare the precision of the estimations of the coefficients. The smaller the standard error, the more reliable the estimated regression coefficient.

Total Residuals, MCR Fitting (Line Plot)

This plot displays the total residuals (all samples and all variables) against increasing number of components in an MCR model.

The size of the residuals is displayed on the scale of the vertical axis. The plot contains one point for each number of components in the model, starting at 2.

The total residuals are a measure of the global fit of the MCR model, equivalent to the total residual variance computed in projection models like PCA.

It may be a good idea to compare the total residuals from an MCR fitting to a PCA fit on the same data (displayed on the plot of Total Residuals, PCA Fitting). Since PCA provides the best possible fit along a set of orthogonal components, the comparison tells you how well the MCR model is performing in terms of fit.

Display the two plots side by side in the Viewer. Check the scale of the vertical axis on each plot (and adjust it if necessary, using **View - Scaling - Min/Max**) before you compare the sizes of the total residuals.

Total Residuals, PCA Fitting (Line Plot)

This plot is available when viewing the results of an MCR model. It displays the total residuals from a PCA model on the same data.

This plot is supposed to be used as a basis for comparison with the Total Residuals, MCR fit (the actual residuals from the MCR model). Since PCA provides the best possible fit along a set of orthogonal components, the comparison tells you how well the MCR model is performing in terms of fit.

Display the two plots side by side in the Viewer. Check the scale of the vertical axis on each plot (and adjust it if necessary, using **View - Scaling - Min/Max**) before you compare the sizes of the total residuals.

Total Variance, X-variables (Line Plot)

This plot gives an indication of how much of the variation in the data is described by the different components.

Total ***residual*** variance is computed as the sum of squares of the residuals for all the variables, divided by the number of degrees of freedom.

Total ***explained*** variance is then computed as:

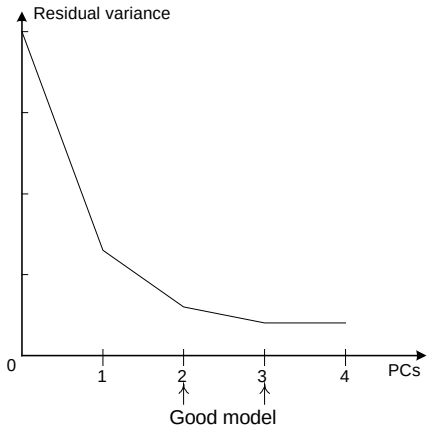
100*(initial variance - residual variance)/(initial variance).

It is the percentage of the original variance in the data which is taken into account by the model.

Both variances can be computed after 0, 1, 2… components have been extracted from the data.

Models with small (close to 0) total residual variance or large (close to 100%) total explained variance explain most of the variation in X; see the example below. Ideally one would like to have simple models, where the residual variance goes to 0 with as few components as possible.

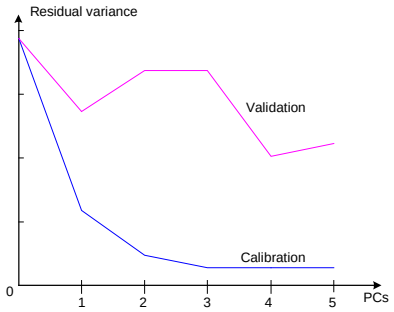
A Total residual variance curve



Calibration variance is based on fitting the calibration data to the model. Validation variance is computed by testing the model on data which was not used to build the model. Compare the two variances: if they differ significantly, there is good reason to question whether either the calibration data or the test data are truly representative. The figure below shows a situation where the residual validation variance is much larger than

the residual calibration variance (or the explained validation variance is much smaller than the explained calibration variance). This means that although the calibration data are well fitted (small residual calibration variances), the model does not describe new data well (large residual validation variance).

Total residual variance curves for Calibration and Validation



Outliers can sometimes cause large residual variance (or small explained variance).

Total Variance, Y-variables (Line Plot)

This plot illustrates how much of the variation in your response(s) is described by each different component.

Total ***residual*** variance is computed as the sum of squares of the residuals for all the variables, divided by the number of degrees of freedom.

Total ***explained*** variance is then computed as:

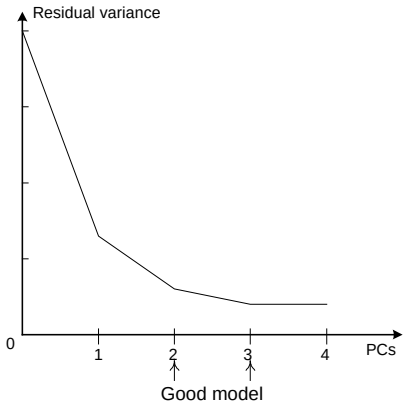
100*(initial variance - residual variance)/(initial variance).

It is the percentage of the original variance in the data which is taken into account by the model.

Both variances can be computed after 0, 1, 2… components have been extracted from the data.

Models with small (close to 0) total residual variance or large (close to 100%) total explained variance explain most of the variation in Y; see the example below for X-variables. Ideally one would like to have simple models, where the residual variance goes to 0 with as few components as possible.

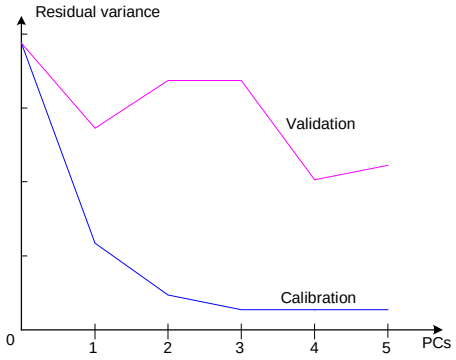
A Total residual variance curve



Calibration variance is based on fitting the calibration data to the model. Validation variance is computed by testing the model on data which was not used to build the model. Compare the two variances: if they differ significantly, there is good reason to question whether either the calibration data or the test data are truly

representative. The figure below shows a situation where the residual validation variance is much larger than the residual calibration variance (or the explained validation variance is much smaller than the explained calibration variance). This means that although the calibration data are well fitted (small residual calibration variances), the model does not describe new data well (large residual validation variance).

Total residual variance curves for Calibration and Validation



Outliers can sometimes be the reason for large residual variance (or small explained variance).

Variable Residuals, MCR Fitting (Line Plot)

This plot displays the residuals for each variable for a given number of components in an MCR model.

The size of the residuals is displayed on the scale of the vertical axis. The plot contains one point for each variable included in the analysis; the variables are listed along the horizontal axis.

The variable residuals are a measure of how well the MCR model takes into account each variable; the better a variable is modeled, the smaller the residual. Variable residuals vary depending on the number of components in the model (displayed in parentheses after the name of the model, at the bottom of the plot). You may tune up or down the number of components for which the residuals are displayed, using the or toolbar buttons.

The size of the residuals tells you about the misfit of the model. It may be a good idea to compare the variable residuals from an MCR fitting to a PCA fit on the same data (displayed on the plot of Variable Residuals, PCA Fitting). Since PCA provides the best possible fit along a set of orthogonal components, the comparison tells you how well the MCR model is performing in terms of fit.

Display the two plots side by side in the Viewer. Check the scale of the vertical axis on each plot to compare the sizes of the residuals.

Variable Residuals, PCA Fitting (Line Plot)

This plot is available when viewing the results of an MCR model. It displays the variable residuals from a PCA model on the same data.

This plot is supposed to be used as a basis for comparison with the Variable Residuals, MCR fit (the actual residuals from the MCR model). Since PCA provides the best possible fit along a set of orthogonal components, the comparison tells you how well the MCR model is performing in terms of fit.

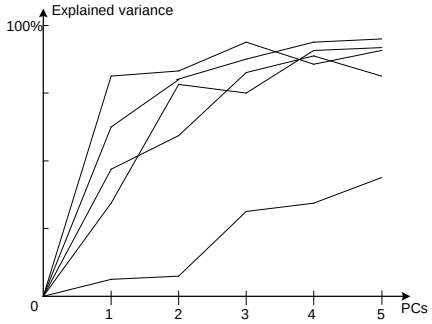
Display the two plots side by side in the Viewer. Check the scale of the vertical axis on each plot to compare the sizes of the residuals.

Variances, Individual X-variables (Line Plot)

This plot shows the ***explained*** or ***residual*** variance for each X-variable when different numbers of components are used in the model. It is used to identify which individual variables are well described by a given model.

X-variables with large explained variance (or small residual variance) for a particular component are explained well by the corresponding model, while those with small explained variance for all (or for at least the first 3-4) components have little relationship to the other X-variables (if this is a PCA model) or little predictive ability (for PCR and PLS models). The figure below shows such a situation, where one X-variable (the lower line) is hardly explained by any of the components.

Explained variances for several individual X-variables



If you find that some variables have much larger residual variance than all the other variables for all components in your model (or for the first 3-4 of them), try rebuilding the model with these variables deleted. This may produce a model which is easier to interpret.

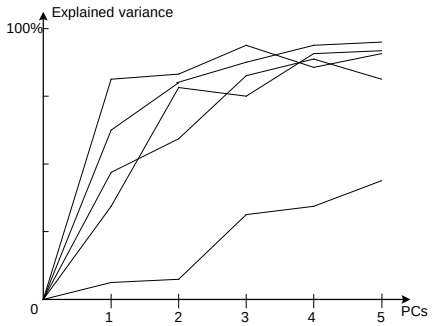
Calibration variance is based on fitting the model to the calibration data. ***Validation*** variance is computed by testing the model on data not used in calibration.

Variances, Individual Y-variables (Line Plot)

This plot shows the ***explained*** or ***residual*** variance for each Y-variable using different numbers of components in the model, and indicates which individual variables are well described by the model.

If a Y-variable has a large explained variance (or small residual variance) for a particular component, it is explained well by the corresponding model. Conversely, Y-variables with small explained variance for all or for the first 3-4 components cannot be predicted from the available X-variables. An example of this is shown below; one variable is poorly explained, even with 5 components.

Explained variances for several individual Y-variables



If some Y-variables have much larger residual variance than the others for all components (or for the first 3-4 of them), you will not be able to predict them correctly. If your purpose is just to interpret variable relationships, you may keep these variables in the model, but remember that they are badly explained. If you intend to make precise predictions, you should recalculate your model without these variables, because the

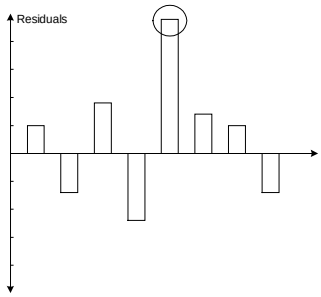
model will not succeed in predicting them anyway. Removing these variables may help the model explain the other Y-variables with fewer components.

Calibration variance is based on fitting the model to the calibration data. ***Validation*** variance is computed by testing the model on new data, not used at the calibration stage. Validation variance is the one which matters most to detect which Y-variables will be predicted correctly.

X-variable Residuals (Line Plot)

This is a plot of residuals for a specified X-variable and component number for all the samples. The plot is useful for detecting outlying sample/variable combinations, as shown below. An outlier can sometimes be modeled by incorporating more such samples. This should, however, be avoided since it will reduce the prediction ability of the model.

Line plot of the variable residuals: one sample is outlying



Whereas the sample residual plot gives information about residuals for all variables for a particular sample, this plot gives information about all possible samples for a particular variable. It is therefore more useful when you want to investigate how one specific variable behaves in all the samples.

X-variable Residuals: Three-way PLS Results

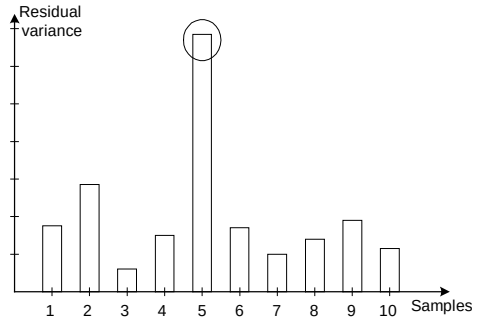
When plotting X-variable residuals from a three-way PLS model, three different cases are encountered. Here follow the details of each case.

- One primary variable selected: a ***matrix*** plot shows the residuals for ***all samples*** x ***all secondary variables***.
- One secondary variable selected: a ***matrix*** plot shows the residuals for ***all samples*** x ***all primary variables***.
- One primary variable and one secondary variable selected: a ***line*** plot shows the residuals for ***all samples***.

X-Variance per Sample (Line Plot)

This plot shows the ***residual*** (or ***explained***) X-variance for all samples, with variable number and number of components fixed. The plot is useful for detecting outlying samples, as shown below. An outlier can sometimes be modeled by incorporating more components. This should be avoided, especially in regression, since it will reduce the predictive power of the model.

An outlying sample has high residual variance



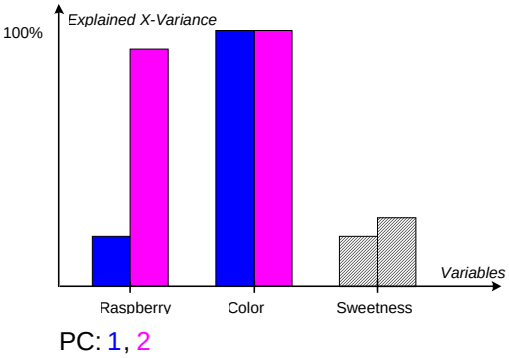
Samples with small residual variance (or large explained variance) for a particular component are well explained by the corresponding model, and vice versa.

X-Variances, One Curve per PC (Line Plot)

This plot displays the variances for all individual X-variables. The horizontal axis shows the X-variables, the vertical axis the variance values. There is one "curve" per PC.

By default, this plot is displayed with a layout as **bars**, and the ***explained variances*** are shown. See the figure below for an illustration.

X-variances for PC1 and PC2, one variable marked



The plot shows which components contribute most to summarizing the variations in each individual variable. For instance, in the example above, PC1 summarizes most of the variations in Color, and PC2 does not add anything to that summary. On the other hand, Raspberry is badly described by PC1, and PC2 is necessary to achieve a good summary.

Use menu option **Edit - Mark - Outliers Only** (or its corresponding shortcut button) if you want the system to mark the badly described variables. For instance, in the example above, variable Sweetness is badly described by a model with 2 components. Try to re-calculate the model with one more component! If you already have many components in your model, badly described variables are either noisy variables (they have little meaningful variations, and can be removed from the analysis), or variables with some data errors.

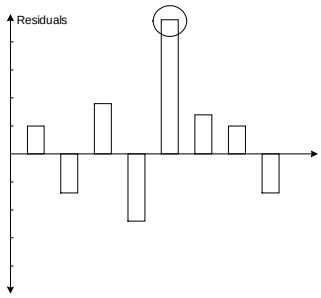
What Should You Do with Your Badly Described X-variables?

First, check their values. You may go back to the outlier plots and search for samples which have outlying values for those variables. If you find an error, correct it. If there is no error, you can re-calculate your model without the marked variables.

Y-variable Residuals (Line Plot)

This is a plot of residuals for a specified Y-variable and component number, for all the samples. The plot is useful for detecting outlying sample or variable combinations, as shown in the figure below. An outlier can sometimes be modeled by incorporating more components. This should be avoided since it will reduce the prediction ability of the model, especially if the outlier is due to an anomaly in your original data (eg. experimental error).

Line plot of the variable residuals: one sample is outlying

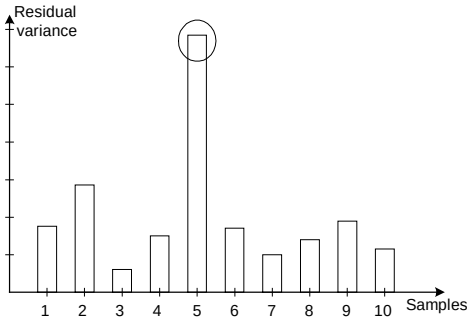


This plot gives information about all possible samples for a particular variable (as opposed to the sample residual plot, which gives information about residuals for all variables for a particular sample) hence it is more useful for studying how a specific variable behaves for all the samples.

Y-Variance Per Sample (Line Plot)

This is a plot of the residual Y-variance for all samples, with fixed variable number and number of components. It is useful for detecting outliers, as shown below. Avoid increasing the number of components in order to model outliers, as this will reduce the predictive power of the model.

An outlying sample has high residual variance



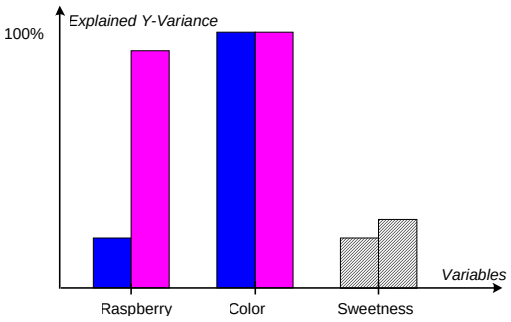
Small residual variance (or large explained variance) indicates that, for a particular number of components, the samples are well explained by the model.

Y-Variances, One Curve per PC (Line Plot)

This plot displays the variances for all individual Y-variables. The horizontal axis shows the Y-variables, the vertical axis the variance values. There is one "curve" per PC.

By default, this plot is displayed with a layout as **bars**, and the ***explained variances*** are shown. See the figure below for an illustration.

Y-variances for PC1 and PC2, one variable marked



PC: 1, 2

The plot shows which components contribute most to summarizing the variations in each individual response variable. For instance, in the example above, PC1 summarizes most of the variations in Color, and PC2 does not add anything to that summary. On the other hand, Raspberry is badly described by PC1, and PC2 is necessary to achieve a good summary.

Use menu option **Edit - Mark - Outliers Only** (or its corresponding shortcut button) if you want the system to mark the badly described variables. For instance, in the example above, variable Sweetness is badly described by a model with 2 components. Try to re-calculate the model with one more component! If you already have many components in your model, badly described response variables are either noisy variables (they have little meaningful variations, and can be removed from the analysis), or variables with some data errors, or responses which cannot be related to the predictors you have chosen to include in the analysis.

What Should You Do with Your Badly Described Y-Variables?

First, check their values. If there is no error, and you have reason to believe that these responses are too noisy, you can re-calculate your model without them. If it seems like some important predictors are missing from your model, you can re-configure the regression calculations and include more predictors, or add interactions and/or squares. If nothing works, you will need to re-think about the whole problem.

2D Scatter Plots

Classification Scores (2D Scatter Plot)

This is a two dimensional scatter plot or map of scores for (PC1,PC2) from a classification. The plot is displayed for one class model at a time. All new samples (the samples you are trying to classify) are shown.

This plot shows how the new samples are projected onto the class model. Members of a particular class are expected to be close to the center of the plot (origo), while non-members should be projected far away from the center.

If you are classifying known samples, this plot helps you detect classification outliers. Look for known members projected far away from the center (false negatives), or known non-members projected close to the center (false positives). There may be errors in the data: check your data and correct them if necessary.

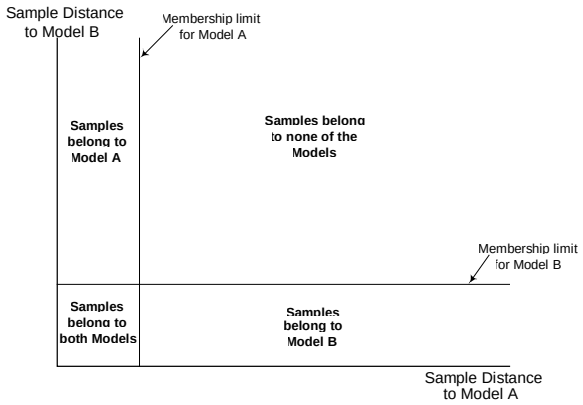
Cooman's Plot (2D Scatter Plot)

This plot shows the orthogonal distances from the new objects to two different classes (models) at the same time. The membership limits (S0) are indicated. Membership limits reflect the significance level used in the classification.

Note: If you select "None" as significance level with the tool when viewing the plot, no membership limits are drawn.

Samples which fall within the membership limit of a class are recognized as members of that class. Different colors denote different types of sample: new samples being classified, calibration samples for the model along the abscissa (A) axis, calibration samples for the model along the ordinate (B) axis, as shown in the figure below.

Cooman's plot


Sample Distance to Model B
Membership limit for Model A
Samples belong to Model A
Samples belong to none of the Models
Membership limit for Model B
Samples belong to both Models
Samples belong to Model B
Sample Distance to Model A

Influence Plot, X-variance (2D Scatter Plot)

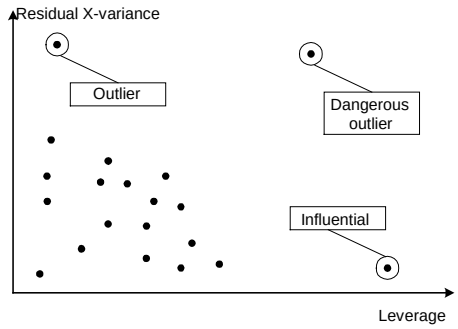
This plot displays the sample residual X-variances against leverages. It is most useful for detecting outliers, influential samples and dangerous outliers.

Samples with high residual variance, i.e. lying to the top of the plot, are likely outliers.

Samples with high leverage, i.e. lying to the right of the plot, are influential; this means that they somehow attract the model so that it describes them better. Influential samples are not necessarily dangerous, if they obey the same model as more "average" samples.

A sample with both high residual variance and high leverage is a dangerous outlier: it is not well described by a model which correctly describes most samples, and it distorts the model so as to be better described, which means that the model then focuses on the difference between that particular sample and the others, instead of describing more general features common to all samples.

Three cases can be detected from the influence plot



Leverages in Designed Data

For designed samples, the leverages should be interpreted differently whether you are running a regression (with the design variables as X-variables) or just describing your responses with PCA.

By construction, the leverage of each sample in the design is known, and these leverages are optimal, i.e. all design samples have the same contribution to the model. So do not bother about the leverages if you are running a regression: the design has cared for it.

However, if you are running a PCA on your response variables, the leverage of each sample is now determined with respect to the response values. Thus some samples may have high leverages, either in an absolute or a relative sense. Such samples are either outliers, or just samples with extreme values for some of the responses.

What Should You Do with an Influential Sample?

The first thing to do is to understand why the sample has a high leverage (and, possibly, a high residual variance). Investigate by looking at your raw data and checking them against your original recordings.

Once you have found an explanation, you are usually in one of the following cases.

Case 1: there is an error in the data. Correct it, or if you cannot find the true value or re-do the experiment which would give you a more valid value, you may replace the erroneous value with “missing”.

Case 2: there is no error, but the sample is different from the others. For instance, it has extreme values for several of your variables. Check whether this sample is “of interest” (e.g. it has the properties you want to achieve, to a higher degree than the other samples), or “not relevant” (e.g. it belongs to another population than the one you want to study). In the former case, you will have to try to generate more samples of the same kind: they are the most interesting ones! In the latter case (and only then), you may remove the high-leverage sample from your model.

Influence Plot, Y-variance (2D Scatter Plot)

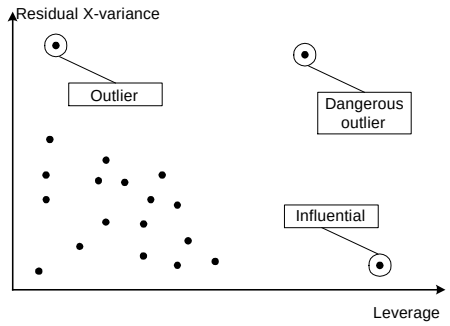
This plot displays the sample residual Y-variances against leverages. It is most useful for detecting outliers, influential samples and dangerous outliers, as shown in the figure below.

Samples with high residual variance, i.e. lying to the top of the plot, are likely outliers, or samples for which the regression model fails to predict Y adequately. To learn more about those samples, study residuals plots (normal probability of residuals, residuals vs. predicted Y values).

Samples with high leverage, i.e. lying to the right of the plot, are influential; this means that they somehow attract the model so that it better describes their X-values. Influential samples are not necessarily dangerous, if they verify the same X-Y relationship as more "average" samples. You can check for that with the X-Y relation outlier plots for several model components.

A sample with both high residual variance and high leverage is a dangerous outlier: it is not well described by a model which correctly describes most samples, and it distorts the model so as to be better described, which means that the model then focuses on the difference between that particular sample and the others, instead of describing more general features common to all samples.

Three cases can be detected from the influence plot



Leverages in Designed Data

By construction, the leverage of each sample in the design is known, and these leverages are optimal, i.e. all design samples have the same contribution to the model. So do not bother about the leverages if you are running a regression on designed samples: the design has cared for it.

What Should You Do with an Influential Sample?

The first thing to do is to understand why the sample has a high leverage (and, possibly, a high residual variance). Investigate, by looking at your raw data, and checking them against your original recordings.

Once you have found an explanation, you are usually in one of the following cases.

Case 1: there is an error in the data. Correct it, or if you cannot find the true value or re-do the experiment which would give you a more valid value, you may replace the erroneous value with "missing".

Case 2: there is no error, but the sample is different from the others. For instance, it has extreme values for several of your variables. Check whether this sample is "of interest" (e.g. it has the properties you want to achieve, to a higher degree than the other samples), or "not relevant" (e.g. it belongs to another population than the one you want to study). In the former case, you will have to try to generate more samples of the same kind: they are the most interesting ones! In the latter case (and only then), you may remove the high-leverage sample from your model.

Loadings for the X-variables (2D Scatter Plot)

A two dimensional scatter plot of X-loadings for two specified components from PCA, PCR, or PLS, this is a good way to detect important variables. The plot is most useful for interpreting component 1 versus component 2, since they represent the largest variations in the X-data (in the case of PCA, as much of the variations as possible for any pair of components).

The plot shows the importance of the different variables for the two components specified. It should preferably be used together with the corresponding score plot. Variables with X-loadings to the right in the loadings plot will be X-variables which usually have high values for samples to the right in the score plot, etc.

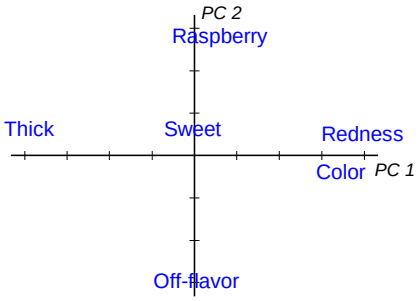
Note: Passified variables are displayed in a different color so as to be easily identified.

Interpretation: X-variables Correlation Structure

Variables close to each other in the loading plot will have a high positive correlation if the two components explain a large portion of the variance of X. The same is true for variables in the same quadrant lying close to a straight line through the origin. Variables in diagonally-opposed quadrants will have a tendency to be negatively correlated. For example, in the figure below, variables Redness and Color have a high positive correlation, and they are negatively correlated to variable Thick. Variables Redness and Off-flavor have

independent variations. Variables Raspberry and Off-flavor are negatively correlated. Variable Sweet cannot be interpreted in this plot, because it is very close to the center.

Loadings of 6 sensory variables along (PC1,PC2)



Note: Variables lying close to the center are poorly explained by the plotted PCs. You cannot interpret them in that plot!

Correlation Loadings Emphasize Variable Correlations

When a PCA, PLS or PCR analysis has been performed and a two dimensional plot of X-loadings is displayed on your screen, you may use the Correlation Loadings option (available from the View menu) to help you discover the structure in the data more clearly.

Correlation loadings are computed for each variable for the displayed Principal Components. In addition, the plot contains two ellipses to help you check how much variance is taken into account. The outer ellipse is the unit-circle and indicates 100% explained variance. The inner ellipse indicates 50% of explained variance.

The importance of individual variables is visualized more clearly in the correlation loading plot compared to the standard loading plot.

Loadings for the Y-variables (2D Scatter Plot)

This is a 2D scatter plot of Y-loadings for two specified components from PCR or PLS and is useful for detecting relevant directions. Like other 2D plots it is particularly useful when interpreting component 1 versus component 2, since these two represent the most important part of the variations in the Y-variables that can be explained by the model.

Note: Passified variables are displayed in a different color so as to be easily identified.

Interpretation: X-Y Relationships in PLS

The plot shows which response variables are well described by the two specified components. Variables with large Y-loadings (either positive or negative) along a component are related to the predictors which have large X-loading weights along the same component.

Therefore, you can interpret X-Y relationships by studying the plot which combines X-loading weights and Y-loadings (see chapter Loading Weights, X-variables, and Loadings, Y-variables (2D Scatter Plot)).

Interpretation: X-Y Relationships in PCR

The plot shows which response variables are well described by the two specified components. Variables with large Y-loadings (either positive or negative) along a component are related to the predictors which have large X-loadings along the same component.

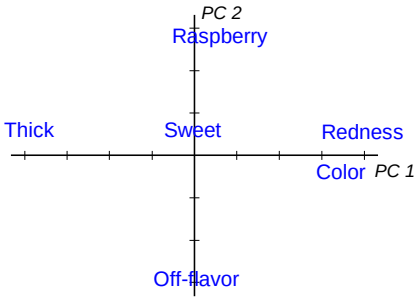
Therefore, you can interpret X-Y relationships by studying the plot which combines X- and Y-loadings (see chapter Loadings for the X- and Y-variables (2D Scatter Plot)).

Interpretation: Y-variables Correlation Structure

Variables close to each other in the loading plot will have a high positive correlation if the two components explain a large portion of the variance of Y. The same is true for variables in the same quadrant lying close to a straight line through the origin. Variables in diagonally-opposed quadrants will have a tendency to be negatively correlated.

For example, in the figure below, variables Redness and Color have a high positive correlation, and they are negatively correlated to variable Thick. Variables Redness and Off-flavor have independent variations. Variables Raspberry and Off-flavor are negatively correlated. Variable Sweet cannot be interpreted in this plot, because it is very close to the center.

Loadings of 6 sensory Y-variables along (PC1,PC2)



Note: Variables lying close to the center are poorly explained by the plotted PCs. You cannot interpret them in that plot!

Correlation Loadings Emphasize Variable Correlations

When a PLS2 or PCR analysis has been performed and a two dimensional plot of Y-loadings is displayed on your screen, you may use the Correlation Loadings option (available from the View menu) to help you discover the structure in your Y-variables more clearly.

Correlation loadings are computed for each variable for the displayed Principal Components. In addition, the plot contains two ellipses to help you check how much variance is taken into account. The outer ellipse is the unit-circle and indicates 100% explained variance. The inner ellipse indicates 50% of explained variance.

The importance of individual variables is visualized more clearly in the correlation loading plot compared to the standard loading plot.

Loadings for the X- and Y-variables (2D Scatter Plot)

This is a 2D scatter plot of X- and Y-loadings for two specified components from PCR. It is used to detect important variables and to understand the relationships between X- and Y-variables. The plot is most useful for interpreting component 1 versus component 2, since these two usually represent the most important part of variation in the data. Note that if you are interested in detecting which X-variables contribute most to predicting the Y-variables, you should preferably choose the plot which combines X-loading weights and Y-loadings.

Note: Passified variables are displayed in a different color so as to be easily identified.

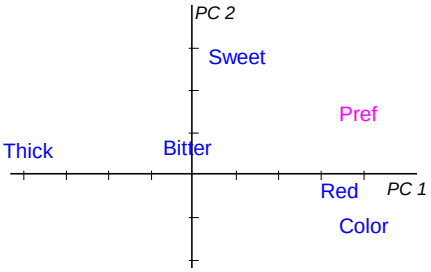
Interpretation: X-Y Relationships

To interpret the relationships between X and Y-variables, start by looking at your response (Y) variables.

- Predictors (X) projected in roughly the same direction from the center as a response, are positively linked to that response. In the example below, predictors Sweet, Red and Color have a positive link with response Pref.

- Predictors projected in the opposite direction have a negative link, as predictor Thick in the example below.
- Predictors projected close to the center, as Bitter in the example below, are not well represented in that plot and cannot be interpreted.

One response (Pref), 5 sensory predictors



Caution!

If your X-variables have been standardized, you should also standardize the Y-variable so that the X- and Y-loadings have the same scale; otherwise the plot may be difficult to interpret.

Correlation Loadings Emphasize Variable Correlations

When a PLS or PCR analysis has been performed and a two dimensional plot of X- and Y-loadings is displayed on your screen, you may use the Correlation Loadings option (available from the View menu) to help you discover the structure in your data more clearly.

Correlation loadings are computed for each variable for the displayed Principal Components. In addition, the plot contains two ellipses to help you check how much variance is taken into account. The outer ellipse is the unit-circle and indicates 100% explained variance. The inner ellipse indicates 50% of explained variance.

The importance of individual variables is visualized more clearly in the correlation loading plot compared to the standard loading plot.

Loading Weights, X-variables (2D Scatter Plot)

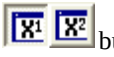
This is a two dimensional scatter plot of X-loading weights for two specified components from a PLS or a tri-PLS analysis.

In PLS, this plot can be useful for detecting which X-variables are most important for predicting Y, although in that case it is better to use the 2D scatter plot of X-loading weights and Y-loadings.

Note: Passified variables are displayed in a different color so as to be easily identified.

X-loading Weights: Three-Way PLS

This is the most important plot of the X-variables in a three-way PLS model. It is especially useful when studied together with a score plot. In that case, interpret the plots in the same way as X-loadings and scores in PCA, PCR or PLS.

Loading weights can be plotted for the Primary or Secondary X-variables. Choose the mode you want to plot in the **2 * 2D Scatter** or **4 * 2D Scatter** sheets of the **Loading Weights** plot dialog, or if the plot is already displayed, use the  buttons to turn off and on one of the modes. The Plot Header tells you which mode is currently plotted (either “X1-loading Weights” or “X2-loading Weights”).

Note: You have to turn off the X-mode currently plotted before you can turn on the other X-mode. This can only be done when Y is also plotted. You may then turn off Y if you are not interested in it.

Read more about:

- How to interpret correlations on a Loading plot, see p.208
- How to interpret scores and loadings together (example of the bi-plot), see p.217

Loading Weights, X-variables, and Loadings, Y-variables (2D Scatter Plot)

This is a 2D scatter plot of X-loading weights and Y-loadings for two specified components from PLS. It shows the importance of the different variables for the two components selected and can thus be used to detect important predictors and understand the relationships between X- and Y-variables. The plot is most useful when interpreting component 1 versus component 2, since these two represent the most important variations in Y.

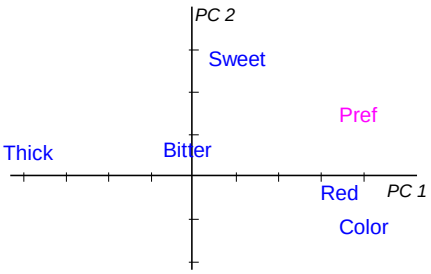
To interpret the relationships between X and Y-variables, start by looking at your response (Y) variables.

Predictors (X) projected in roughly the same direction from the center as a response, are positively linked to that response. In the example below, predictors Sweet, Red and Color have a positive link with response Pref.

Predictors projected in the opposite direction have a negative link, as predictor Thick in the example below.

Predictors projected close to the center, as Bitter in the example above, are not well represented in that plot and cannot be interpreted.

One response (Pref), 5 sensory predictors



Note: Passified variables are displayed in a different color so as to be easily identified.

Scaling the Variables and the Plot

Here are two important details you should watch if you want to make sure that you are interpreting your plot correctly.

1- For PLS1, if your ***X-variables*** have been **standardized**, you should also standardize the ***Y-variable*** so that the X-loading weights and Y-loadings have the same scale; otherwise the plot may be difficult to interpret.

2- Make sure that the ***two axes*** of the plot have **consistent scales**, so that a unit of 1 horizontally is displayed with the same size as a unit of 1 vertically. This is the necessary condition for interpreting directions correctly.

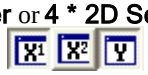
Interpretation for more than 2 Components

If your PLS model has more than 2 useful components, this plot is still interesting, because it shows the correlations among predictors, among responses, and between predictors and responses, along each component. However, you will get a better summary of the relationships between X and Y by looking at the regression coefficients, which take into account all useful components together.

X-loading Weights and Y-loadings: Three-Way PLS

In a three-way PLS model, X- and Y-variables both have a set of loading weights (sometimes also just called **weights**). However, the plot is still referred to as resp. "X1-loading Weights and Y-loadings" or "X2-loading Weights and Y-loadings".

The plot reveals relationships between X- and Y-variables in the same way as X-loading Weights and Y-loadings in PLS.

X-loading weights are plotted either for the Primary or Secondary X-variables. Choose the mode you want to plot in the **2 * 2D Scatter** or **4 * 2D Scatter** sheets of the **Loading Weights** plot dialog, or if the plot is already displayed, use the  buttons to turn off and on one of the modes. The Plot Header tells you which mode is currently plotted (either "X1-loading Weights and Y-loadings" or "X2-loading Weights and Y-loadings").

Note: You have to turn off the X-mode currently plotted before you can turn on the other X-mode. This can only be done when Y is also plotted.

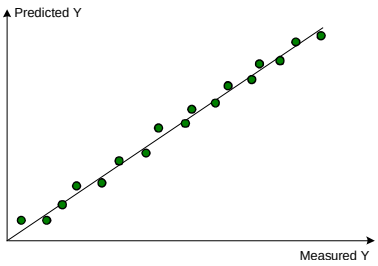
Predicted vs. Measured (2D Scatter Plot)

The predicted Y-value from the model is plotted against the measured Y-value. This is a good way to check the quality of the regression model. If the model gives a good fit, the plot will show points close to a straight line through the origin and with slope equal to 1. Turn on **Plot Statistics** (using the **View** menu) to check the slope and offset, and RMSEP/RMSEC.

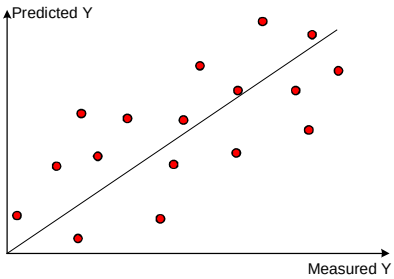
The figures below show two different situations: one indicating a good fit, the other a poor fit of the model.

Predicted vs. Measured shows how well the model fits

Good fit:

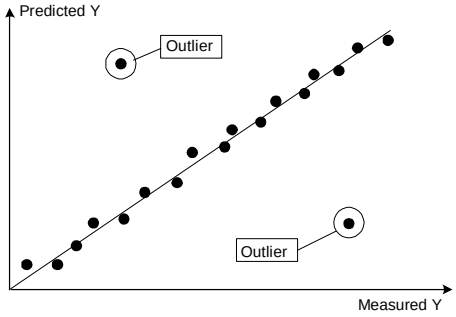


Bad fit:



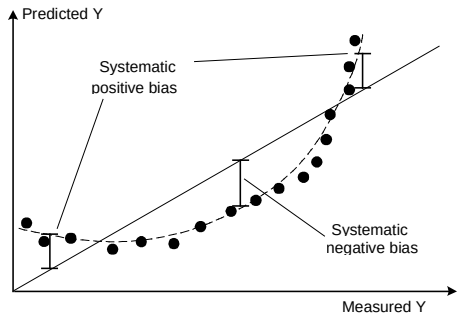
You may also see cases where the majority of the samples lie close to the line while a few of them are further away. This may indicate good fit of the model to the majority of the data, but with a few outliers present (see the figure below).

Detecting outliers on a Predicted vs. Measured plot



In other cases, there may be a non-linear relationship between the X- and Y-variables, so that the predictions do not have the same level of accuracy over the whole range of variation of Y. In such cases, the plot may look like the one shown below. Such non-linearities should be corrected if possible (for instance by a suitable transformation), because otherwise there will be a systematic bias in the predictions depending on the range of the sample.

Predicted vs. Measured shows a non-linear relationship



Predicted vs. Reference (2D Scatter Plot)

This is a plot of predicted Y-values versus the true (measured) reference Y-values. You can use it to check whether the model predicts new samples well. Ideally the predicted values should be equal to the reference values.

Note that this plot is built in the same way as the Predicted vs. Measured plot used during calibration. You can also turn on **Plot Statistics** (use the **View** menu) to display the slope and offset of the regression line, as well as the true value of the RMSEP for your predicted values.

Projected Influence Plot (3 x 2D Scatter Plots)

This is the projected view of a 3D influence plot. In addition to the original 3D plot, you can see the following:

- 2D influence plot with X-residual variance;
- 2D influence plot with Y-residual variance;
- X-residual variance vs. Y-residual variance.

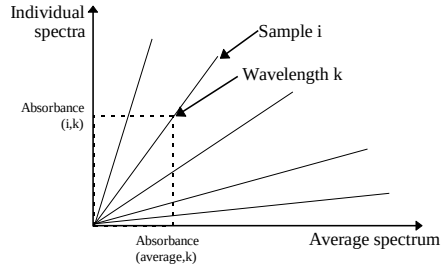
Scatter Effects (2D Scatter Plot)

This plot shows each sample plotted against the average sample. Scatter effects appear as differences in slope and/or offset between the lines in the plot. Differences in the slope are caused by multiplicative scatter effects. Offset error is due to additive effects.

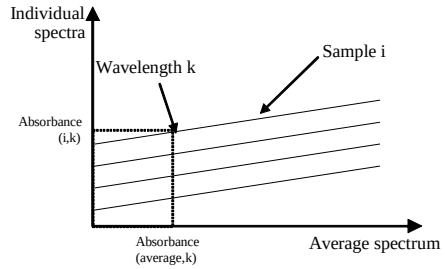
Applying Multiplicative Scatter Correction will improve your model if you detect these scatter effects in your data table. The examples below show what to look for.

Two cases of scatter effects

Multiplicative Scatter Effect



Additive Scatter Effect



Read more about:

- How Multiplicative Scatter Correction works, see p. **Feil! Bokmerke er ikke definert.**
- How to apply Multiplicative Scatter Correction, see p. 87

Scores (2D Scatter Plot)

This is a two dimensional scatter plot (or map) of scores for two specified components (PCs) from PCA, PCR, or PLS. The plot gives information about patterns in the samples. The score plot for (PC1,PC2) is especially useful, since these two components summarize more variation in the data than any other pair of components.

The closer the samples are in the score plot, the more similar they are with respect to the two components concerned. Conversely, samples far away from each other are different from each other.

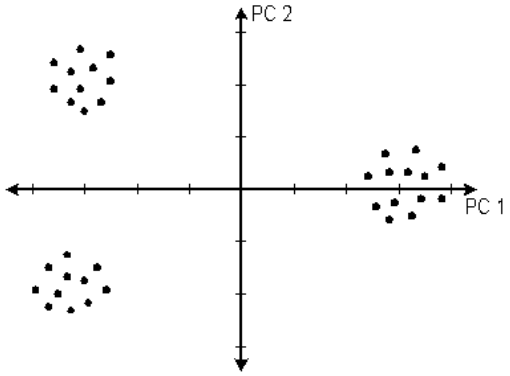
The plot can be used to interpret differences and similarities among samples. Look at the present plot together with the corresponding loading plot, for the same two components. This can help you determine which variables are responsible for differences between samples. For example, samples to the right of the score plot will usually have a large value for variables to the right of the loading plot, and a small value for variables to the left of the loading plot.

Here are some things to look for in the 2D score plot.

Finding Groups in a Score Plot

Is there any indication of ***clustering*** in the set of samples? The figure below shows a situation with three distinct clusters. Samples within a cluster are similar.

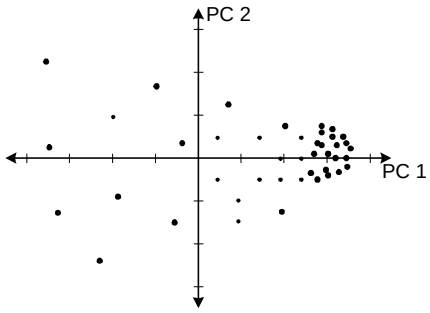
Three groups of samples



Studying Sample Distribution in a Score Plot

Are the samples ***evenly spread*** over the whole region, or is there any accumulation of samples at one end? The figure below shows a typical fan-shaped layout, with most samples accumulated to the right of the plot, then progressively spreading more and more. This means that the variables responsible for the major variations are asymmetrically distributed. If you encounter such a situation, study the **distributions** of those variables (histograms), and use an appropriate **transformation** (most often a logarithm).

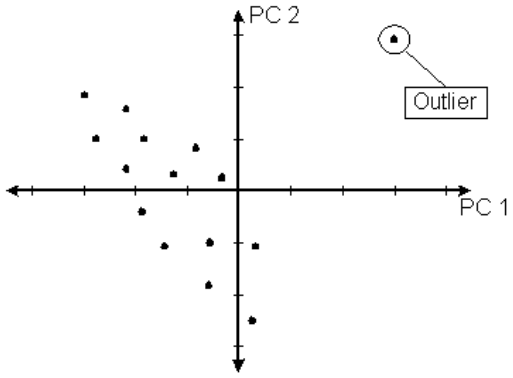
Asymmetrical distribution of the samples on a score plot



Detecting Outliers in a Score Plot

Are some samples very ***different from the rest***? This can indicate that they are **outliers**, as shown in the figure below. Outliers should be investigated: there may have been errors in data collection or transcription, or those samples may have to be removed if they do not belong to the population of interest.

An outlier sticks out of the major group of samples



How Representative Is the Picture?

Check how much of the total variation each of the components explains. This is displayed in parentheses at the bottom of the plot. If the sum of the explained variances for the 2 components is large (for instance 70-80%), the plot shows a large portion of the information in the data, so you can interpret the relationships with a high degree of certainty. On the other hand if it is smaller, you may need to study more components or consider a transformation, or there may simply be little meaningful information in your data.

Scores and Loadings (Bi-plot)

This is a two dimensional scatter plot or map of scores for two specified components (PCs), with the X-loadings displayed on the same plot. It is called a bi-plot. It enables you to interpret sample properties and variable relationships simultaneously.

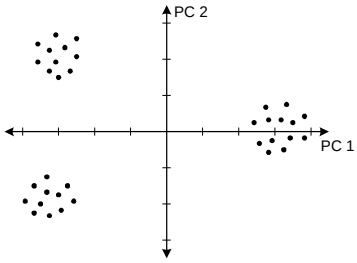
Scores

The closer two samples are in the score plot, the more similar they are with respect to the two components concerned. Conversely, samples far away from each other are different from each other.

Here are a few things to look for in the score plot

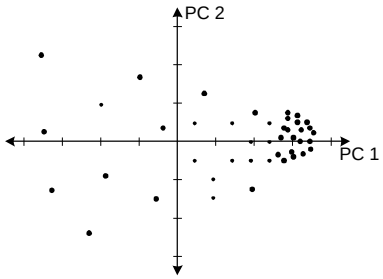
1- Is there any indication of ***clustering*** in the set of samples? The figure below shows a situation with three distinct clusters. Samples within a cluster are similar.

Three groups of samples



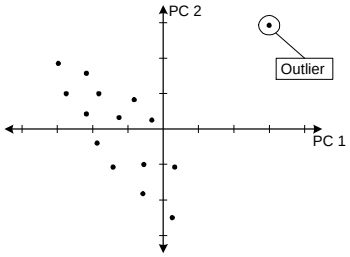
2- Are the samples ***evenly spread*** over the whole region, or is there any accumulation of samples at one end? The figure below shows a typical fan-shaped layout, with most samples accumulated to the right of the plot, then progressively spreading more and more. This means that the variables responsible for the major variations are asymmetrically distributed. If you encounter such a situation, study the **distributions** of those variables (histograms), and use an appropriate **transformation** (most often a logarithm).

Asymmetrical distribution of the samples on a score plot



3- Are some samples very ***different from the rest***? This can indicate that they are **outliers**, as shown in the figure below. Outliers should be investigated: there may have been errors in data collection or transcription, or those samples may have to be removed if they do not belong to the population of interest.

An outlier sticks out of the major group of samples



Loadings

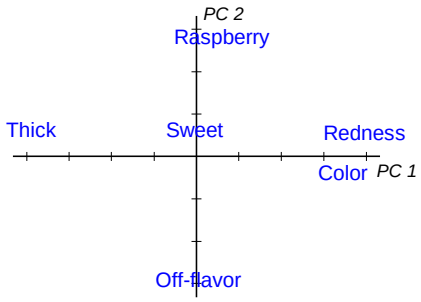
The plot shows the importance of the different variables for the two components specified. Variables with loadings to the right in the loadings plot will be variables which usually have high values for samples to the right in the score plot, etc.

Note: Passified variables are displayed in a different color so as to be easily identified.

Interpret variable projections on the loading plot

Variables close to each other in the loading plot will have a high positive correlation if the two components explain a large portion of the variance of X. The same is true for variables in the same quadrant lying close to a straight line through the origin. Variables in diagonally-opposed quadrants will have a tendency to be negatively correlated. For example, in the figure below, variables Redness and Color have a high positive correlation, and they are negatively correlated to variable Thick. Variables Redness and Off-flavor have independent variations. Variables Raspberry and Off-flavor are negatively correlated. Variable Sweet cannot be interpreted in this plot, because it is very close to the center.

Loadings of 6 sensory variables along (PC1,PC2)



Scores and Loadings Together

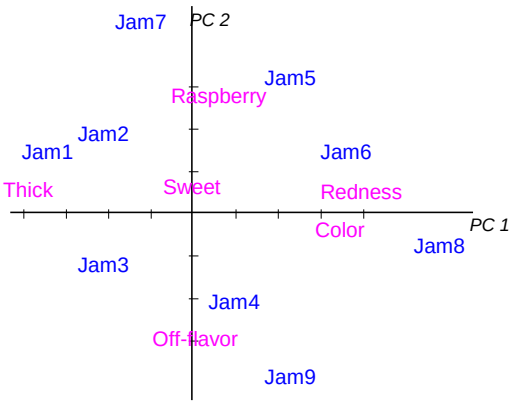
The plot can be used to interpret ***sample properties***. Look for variables projected far away from the center. Samples lying in an extreme position in the same direction as a given variable have large values for that variable; samples lying in the opposite direction have low values.

For instance, in the figure below, Jam8 is the most colorful, while Jam9 has the highest off-flavor (and probably lowest Raspberry taste). Jam9 is very different from Jam7: Jam7 has highest Raspberry taste and lowest off-flavor, otherwise those two jams do not differ much in color and thickness.

Jam5 has high Raspberry taste, and is rather colorful. Jam1, Jam2 and Jam3 are thick, and have little color.

The jams cannot be compared with respect to sweetness, because variable Sweet is projected close to the center.

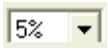
Bi-plot for 8 jam samples and 6 sensory properties



Note: Passified variables are displayed in a different color so as to be easily identified.

Si vs. Hi (2D Scatter Plot)

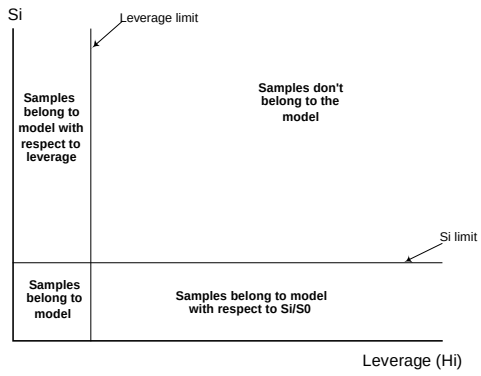
The Si vs. Hi plot shows the two limits used for classification. Si is the distance from the new sample to the model (square root of the residual variance) and Hi is the leverage (distance from the projected sample to the model center).

Note: If you select "None" as significance level with the 

tool when viewing the plot, no membership limits are drawn.

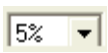
Samples falling within both limits for a class are recognized as members of that class. The level of the limits is governed by the significance level used in the classification.

Membership limits on the Si vs. Hi plot


Si
Leverage limit
Samples belong to model with respect to leverage
Samples don't belong to the model
Si limit
Samples belong to model
Samples belong to model with respect to Si/S0
Leverage (Hi)

Si/S0 vs. Hi (2D Scatter Plot)

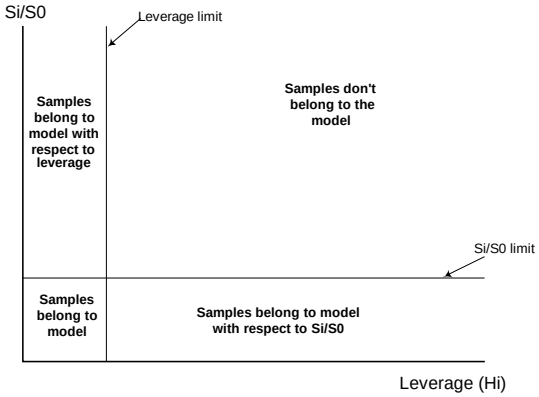
The Si/S0 vs. Hi plot shows the two limits used for classification: the relative distance from the new sample to the model (residual standard deviation) and the leverage (distance from the new sample to the model center).

Note: If you select "None" as significance level with the 

tool when viewing the plot, no membership limits are drawn.

Samples which fall within both limits for a particular class are said to belong to that class. The level of the limits is governed by the significance level used in the classification.

Membership limits on the Si/S0 vs. Hi plot


Si/S0
Leverage limit
Samples belong to model with respect to leverage
Samples don't belong to the model
Si/S0 limit
Samples belong to model
Samples belong to model with respect to Si/S0
Leverage (Hi)

X-Y Relation Outliers (2D Scatter Plot)

This plot visualizes the regression relation along a particular component of the PLS model. It shows the t-scores as abscissa and the u-scores as ordinate. In other words, it shows the relationship between the projection of your samples in the X-space (horizontal axis) and the projection of your samples in the Y-space (vertical axis).

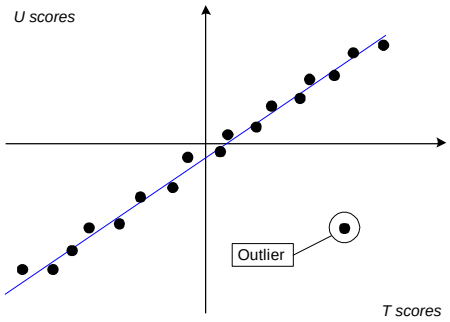
Note: The X-Y relation outlier plot for PC1 is exactly the same as Predicted vs. Measured for PC1.

This summary can be used for two purposes.

Detecting Outliers

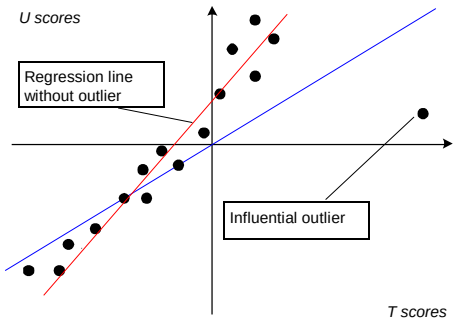
A sample may be outlying according to the X-variables only, or to the Y-variables only, or to both. It may also not have extreme or outlying values for either separate set of variables, but become an outlier when you consider the (X,Y) relationship. In the X-Y Relation Outlier plot, such a sample sticks out as being far away from the relation defined by the other samples, as shown in the figure below. Check your data: there may be a data transcription error for that sample.

A simple X-Y outlier



If a sample sticks out in such a way that it is projected far away from the center along the model component, we have an influential outlier (see the figure below). Such samples are dangerous to the model: they change the orientation of the component. Check your data. If there is no data transcription error for that sample, investigate more and decide whether it belongs to another population. If so, you may remove that sample (mark it and recalculate the model without the marked sample). If not, you will have to gather more samples of the same kind, in order to make your data more balanced.

An influential outlier

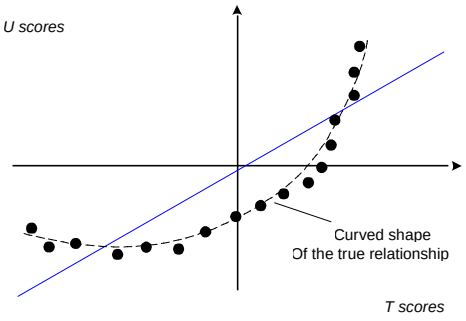


Studying The Shape of the X-Y Relationship

One of the underlying assumptions of PLS is that the relationship between the X- and Y-variables is essentially linear. A strong deviation from that assumption may result in unnecessarily high calibration or prediction errors. It will also make the prediction error unevenly spread over the range of variation of the response. Thus it is important to detect non-linearities in the X-Y relation (especially if they occur in the first model components), and try to correct them.

An exponential-like curvature, as in the figure below, may appear when one or several responses have a skewed (asymmetric) distribution. A logarithmic transformation of those variables may improve the quality of the model.

Non-linear relationship between X and Y



A sigmoid-shaped curvature may indicate that there are interactions between the predictors. Adding cross-term to the model may improve it.

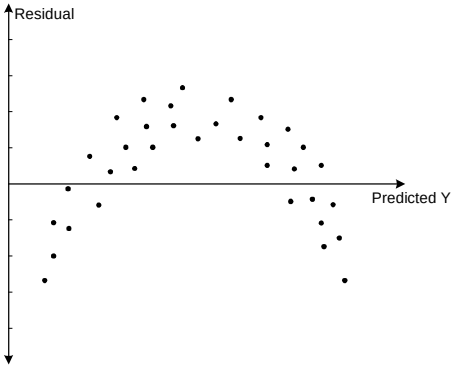
Sample groups may indicate the need for separate modeling of each subgroup.

Y-Residuals vs. Predicted Y (2D Scatter Plot)

This is a plot of Y-residuals against predicted Y values. If the model adequately predicts variations in Y, any residual variations should be due to noise only, which means that the residuals should be randomly distributed. If this is not the case, the model is not completely satisfactory, and appropriate action should be taken.

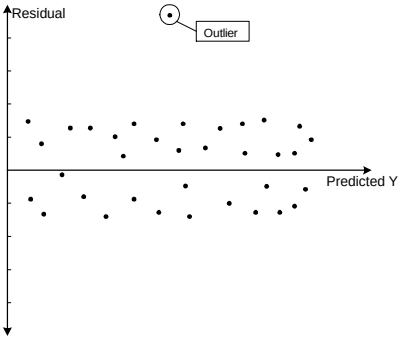
If strong systematic structures (e.g. curved patterns) are observed, this can be an indication of lack of fit of the regression model. The figure below shows a situation which strongly indicates lack of fit of the model. This may be corrected by transforming the Y variable.

Structure in the residuals: you need a transformation



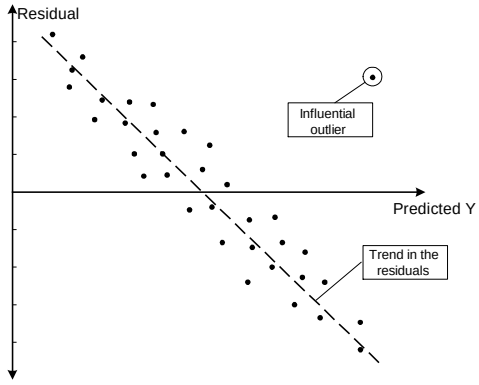
The presence of an outlier is shown in the example below. The outlying sample has a much larger residual than the others; however, it does not seem to disturb the model to a large extent.

A simple outlier has a large residual



The figure below shows the case of an influential outlier: not only does it have a large residual, it also attracts the whole model so that the remaining residuals show a very clear trend. Such samples should usually be excluded from the analysis, unless there is an error in the data or some data transformation can correct for the phenomenon.

An influential outlier changes the structure of the residuals

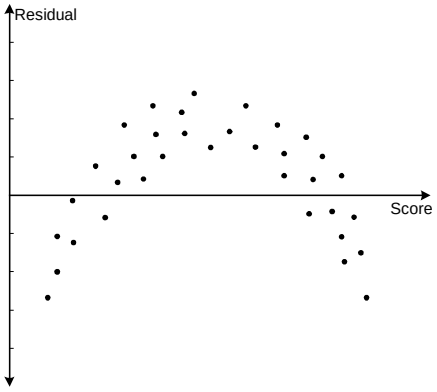


Small residuals (compared to the variance of Y) which are randomly distributed indicate adequate models.

Y-Residuals vs. Scores (2D Scatter Plot)

This is a plot of Y-residuals versus component scores. Clearly visible structures are an indication of lack of fit of the regression model. The figure below shows such a situation, with a strong nonlinear structure of the residuals indicating lack of fit. We can say that there is a lack of fit in the direction (in the multidimensional space) defined by the selected component. Small residuals (compared to the variance of Y) which are randomly distributed indicate adequate models.

Structure in the residuals: you need a transformation



3D Scatter Plots

Influence Plot, X- and Y-variance (3D Scatter Plot)

This is a plot of the residual X- and Y-variances versus leverages. Look for samples with a high leverage and high residual X- or Y-variance.

To study such samples in more detail, we recommend that you mark them and then plot X-Y relation outliers for several model components. This way you will detect whether they have an influence on the shape of the X-Y relationship, in which case they would be dangerous outliers.

The plot is usually easier to read in its "projected" version. See Projected Influence Plot (3 x 2D Scatter Plots) for more details.

Loadings for the X-variables (3D Scatter Plot)

This is a three-dimensional scatter plot of X-loadings for three specified components from PCA, PCR, or PLS. The plot is most useful for interpreting directions, in connection to a 3D score plot. Otherwise we would recommend that you use line- or 2D loading plots.

Note: Passified variables are displayed in a different color so as to be easily identified.

Loadings for the X- and Y-variables (3D Scatter Plot)

This is a three dimensional scatter plot of X- and Y-loadings for three specified components from PCR or PLS. The plot is most useful for interpreting directions, in connection to a 3D score plot. Otherwise we would recommend that you use line- or 2D loading plots.

Note: Passified variables are displayed in a different color so as to be easily identified.

Loadings for the Y-variables (3D Scatter Plot)

This is a three dimensional scatter plot of Y-loadings for three specified components from PLS. The plot is most useful for interpreting directions, in connection to a 3D score plot. Otherwise we would recommend that you use line- or 2D loading plots.

Note: Passified variables are displayed in a different color so as to be easily identified.

Loading Weights, X-variables (3D Scatter Plot)

This is a three dimensional scatter plot of X-loading weights for three specified components from PLS; this plot may be difficult to interpret, both because it is three-dimensional and because it does not include the Y-loadings. Thus we would usually recommend that you use the 2D scatter plot of X-loading weights and Y-loadings instead.

Note: Passified variables are displayed in a different color so as to be easily identified.

Loading Weights, X- variables, and Loadings, Y-variables (3D Scatter Plot)

This is a three dimensional scatter plot of X-loading weights and Y-loadings for three specified components from PLS, showing the importance of the different X-variables for the prediction of Y. Since such 3D plots are often difficult to read, we would usually recommend that you use the 2D scatter plot of X-loading weights and Y-loadings instead.

Note: Passified variables are displayed in a different color so as to be easily identified.

Scores (3D Scatter Plot)

This is a 3D scatter plot or map of the scores for three specified components from PCA, PCR, or PLS. The plot gives information about patterns in the samples and is most useful when interpreting components 1, 2 and 3, since these components summarize most of the variation in the data. It is usually easier to look at 2D score plots but if you need three components to describe enough variation in the data, the 3D plot is a practical alternative.

Like with the 2D plot, the closer the samples are in the 3D score plot, the more similar they are with respect to the three components.

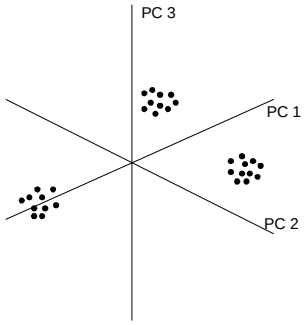
The 3D plot can be used to interpret differences and similarities among samples. Look at the score plot and the corresponding loadings plot, for the same three components. Together they can be used to determine which variables are responsible for differences between samples. Samples with high scores along the first component usually have a large values for variables with high loadings along the first component, etc.

Here are a few patterns to look for in a score plot.

Finding Groups in a Score Plot

Do the samples show any tendency towards clustering? A plot with three distinct clusters is shown below. Samples within the same cluster are similar to each other.

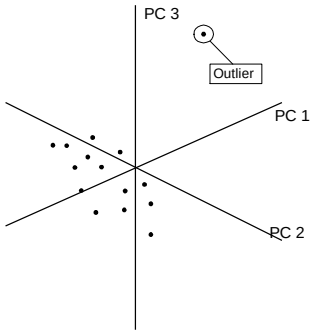
Three groups of samples appear on the score plot



Detecting Outliers in a Score Plot

Are one or more samples very different from the rest? If so, this can indicate that they are outliers. A situation with an outlying sample is given in the figure below. Outliers may have to be removed.

An outlier sticks out of the main group of samples



Check how much of the total variation is explained by each component (these numbers are displayed at the bottom of the plot). If it is large, the plot shows a significant portion of the information in your data and you can use it to interpret relationships with a high degree of certainty. If the explained variation is smaller, you may need to study more components, consider a transformation, or there may be little information in the original data.

Matrix Plots

Leverages (Matrix Plot)

This is a matrix plot of leverages for all samples and all model components. It is a useful plot for studying how the influence of each sample evolves with the number of components in the model.

Mean (Matrix Plot)

For each analyzed variable, the average over all samples in each group is displayed. The groups correspond to the levels of all leveled variables (design or category variables) contained in the data set.

This plot can be useful to detect main effects of variables, by comparing the averages between various levels of the same leveled variable.

Regression Coefficients (Matrix Plot)

Regression coefficients summarize the relationship between all predictors and a given response. For PCR and PLS, the regression coefficients can be computed for any number of components. The regression coefficients for 5 PCs, for example, summarize the relationship between the predictors and the response, as it is approximated by a model with 5 components.

Note: What follows applies to a matrix plot of regression coefficients in general. To read about specific features related to three-way PLS results, look up the Details section below.

This plot shows an overview of the regression coefficients for all response variables (Y), and all predictor variables (X). It is displayed for a model with a particular number of components. You can choose a layout as **bars**, or as **map**.

The regression coefficients matrix plot is available in two options: weighted coefficients (BW), or raw coefficients (B).

Note: The weighted coefficients (BW) and raw coefficients (B) are identical if no weights where applied on your variables.

If you have weighted your predictor variables with 1/Sdev (standardization), the ***weighted*** regression coefficients (BW) take these weights into account. Since all predictors are brought back to the same scale, the coefficients show the relative importance of those variables in the model.

- Predictors with a large weighted coefficient play an important role in the regression model; a positive coefficient shows a positive link with the response, and a negative coefficient shows a negative link.
- Predictors with a small weighted coefficient are negligible. You can recalculate the model without those variables.

The ***raw*** regression coefficients are those that may be used to write the model equation in original units:

Y = B0 + B1 * X-variable1 + B2 * X-variable2 + …

Since the predictors are kept in their original scales, the coefficients do not reflect the relative importance of the X-variables in the model.

The raw coefficients do **not** reflect the importance of the X-variables in the model, because the sizes of these coefficients depend on the range of variation (and indirectly, on the original units) of the X-variables.

- A predictor with a small raw coefficient does not necessarily indicate an unimportant variable
- A predictor with a large raw coefficient does not necessarily indicate an important variable.

Matrix Plot of Regression Coefficients: Three-Way PLS

In a three-way PLS model, Primary and Secondary X-variables both have a set of regression coefficients (one for each Y-variable).

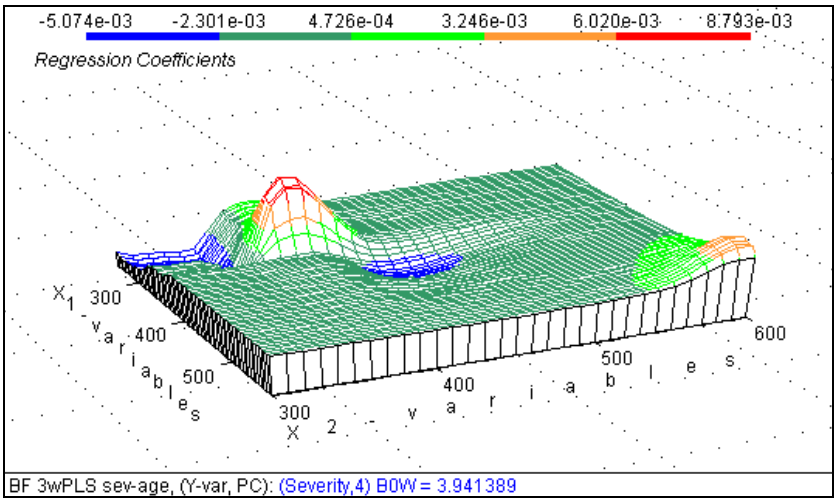
Thus, if you have several Y-variables, there are three relevant ways to study the regression coefficients as a matrix:

- X1 vs X2 (for a selected Response Y)
- X1 vs Y (for a selected Secondary X-variable X2)
- X2 vs Y (for a selected Primary X-variable X1)

If you have only one response, the first plot is relevant while the other two can be replaced by a Line plot of the regression coefficients.

The matrix plot of X1- vs X2-regression coefficients gives you a graphical overview of the regions in your 3-D arrays which are important for a given response. In the example below, you can see that most of the information relevant to the prediction of response “Severity” is concentrated around X1= 250-400 and X2= 300-450, with an additional interesting spot around X1=550 and X2=600.

X1 vs X2 Matrix plot of Regression Coefficients for response Severity



If you have several responses, use the X1 vs Y and X2 vs Y plots to get an overview of one mode with respect to all responses simultaneously. This will allow you to answer questions such as:

- Is there a region of mode 1 (resp. 2) which is important for several responses?

- Is the relationship between X1 and Y the same for all responses?

- Is there a region of mode 1 (resp. 2) which does not play any role for any of the responses? If so, it may be removed from future models.

Response Surface (Matrix Plot)

This plot is used to find the settings of the design variables which give an optimal response value, and to study the general shape of the response surface fitted by the Response Surface model or the Regression model. It shows one response variable at a time. For PCR or PLS models, it uses a certain number of components. Check that this is the optimal number of components before interpreting your results!

This plot can appear in various layouts. The most relevant are:

- Contour plot;
- Landscape plot.

Interpretation: Contour Plot

Look at this plot if you want a map which tells you how to reach your goal. The plot has two axes: two predictor variables are studied over their range of variation, the remaining ones are kept constant. The constant levels are indicated in the **Plot ID** at the bottom.

The response values are displayed as contour lines, i.e. lines which show where the response variable has the same predicted value. Clicking on a line, or on any spot within the map, will tell you the predicted response value for that point, and the coordinates of the point (i.e. the settings of the two predictor variables giving that particular response value).

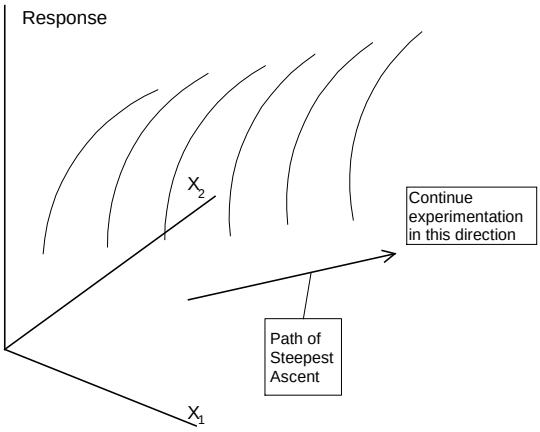
If you want to interpret several responses together, print out their contour plots on color transparencies and superimpose the maps.

Interpretation: Landscape Plot

Look at this plot if you want to study the 3D shape of your response surface. Here it is obvious whether you have a maximum, a minimum or a saddle point.

This plot, however, does not tell you precisely how the optimum you are looking for can be achieved.

Response surface plot, with Landscape layout



Sample and Variable Residuals, X-variables (Matrix Plot)

This is a plot of the residuals for all X-variables and samples for a specified component number. It can be used to detect outlying (sample*variable) combinations.

An outlier can be recognized by looking for high residuals. Sometimes outliers can be modeled by incorporating more components in the model. This should be avoided as it will reduce the prediction ability of the model.

Sample and Variable Residuals, Y-variables (Matrix Plot)

This is a plot of the residuals for all Y-variables and samples for a specified component number. The plot is useful for detecting outlying (sample*variable) combinations.

High residuals indicate an outlier. Incorporating more components can sometimes model outliers; you should avoid doing so since it will reduce the prediction ability of your model.

Standard Deviation (Matrix Plot)

For each variable, the standard deviation (square root of the variance) is displayed over each group. The groups correspond to the levels of all leveled variables (design or category variables) contained in the data set.

Cross-Correlation (Matrix Plot)

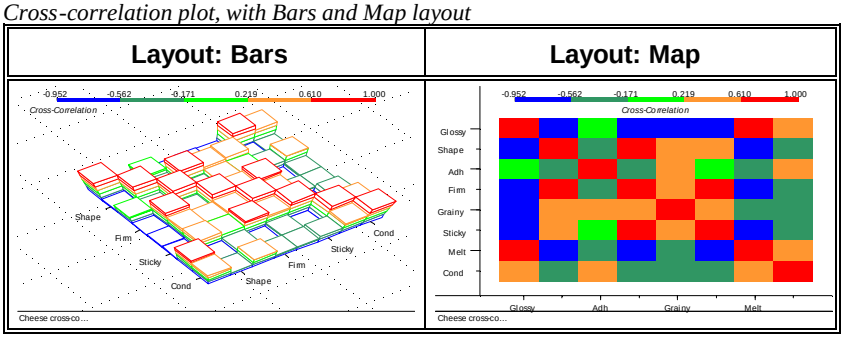
This plot shows the cross-correlations between all variables included in a Statistics analysis.

The matrix is symmetrical (the correlation between A and B is the same as between B and A) and its diagonal contains only values of 1, since the correlation between a variable and itself is 1.

All other values are between -1 and +1. A large positive value (as shown in red on the figure below) indicates that the corresponding two variables have a tendency to increase simultaneously. A large negative value (as shown in blue on the figure below) indicates that when the first variable increases, the other often decreases. A correlation close to 0 (light green on the figure below)indicates that the two variables vary independently from each other.

The best layouts for studying cross-correlations are “bars” (used as default) or “map”.

Cross-correlation plot, with Bars and Map layout



Note:

Be careful when interpreting the color scale of the plot; not all data sets have correlations varying from -1 to +1. The highest value will always be +1 (diagonal), but the lowest may not even be below zero! This may happen for instance if you are studying several measurements that all capture more or less the same phenomenon, e.g. texture or light absorbance in a narrow range.

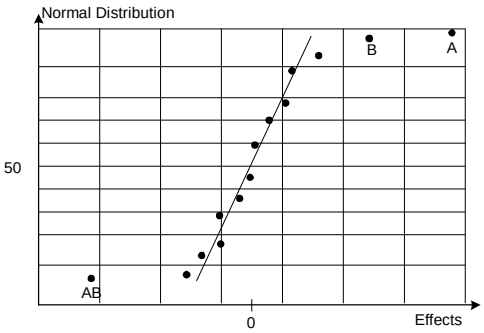
Look at the values on the color scale before jumping to conclusions!

Normal Probability Plots

Effects (Normal Probability Plot)

This is a normal probability plot of all the effects included in an Analysis of Effects model. Effects in the upper right or lower left of the plot deviating from a fictitious straight line going through the medium effects are potentially significant. The figure below shows such an example where A, B, and AB are potentially significant. More specific results about significance can be obtained from other plots, for instance the line plot of individual effects with p-values, or the effects table.

Two positive and one negative effect are sticking out



You may manually draw a line on the plot with menu option **Edit - Insert Draw Item - Line**.

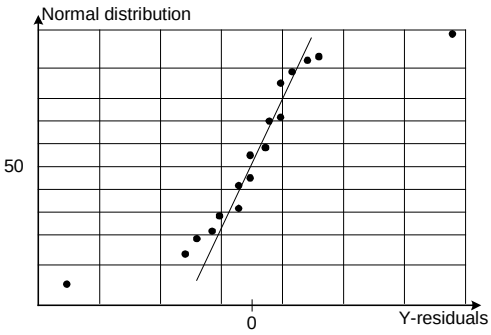
Y-residuals (Normal Probability Plot)

This plot displays the cumulative distribution of the Y-residuals with a special scale, so that normally distributed values should appear along a straight line. The plot shows all residuals for one particular Y-variable (look for its name in the plot ID). There is one point per sample.

If the model explains the complete structure present in your data, the residuals should be randomly distributed - and usually, normally distributed as well. So if all your residuals are along a straight line, it means that your model explains everything which can be explained in the variations of the variables you are trying to predict.

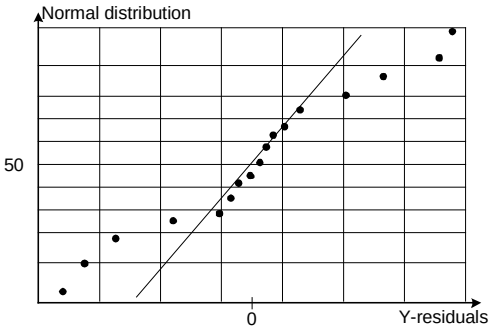
If most of your residuals are normally distributed, and one or two stick out, these particular samples are outliers. This is shown in the figure below. If you have outliers, mark them and check your data.

Two outliers are sticking out



If the plot shows a strong deviation from a straight line, the residuals are not normally distributed, as in the figure below. In some cases - but not always - this can indicate lack of fit of the model. However it can also be an indication that the error terms are simply not normally distributed..

The residuals have a regular but non-normal distribution



You may manually draw a line on the plot with menu option **Edit - Insert Draw Item - Line**.

Table Plots

ANOVA Table (Table Plot)

The ANOVA table contains degrees of freedom, sums of squares, mean squares, F-values and p-values for all sources of variation included in the model.

The Multiple Correlation coefficient and the R-square are also presented above the main table. A value close to 1 indicates a good fit, while a value close to 0 indicates a poor fit.

For Response surface analyses, a Model check and a Lack of fit test are displayed after the Variables part of the ANOVA table. The table may also include a significance test for the intercept, and the coordinates of max/min/saddle points.

First Section: Summary

The first part of the ANOVA table is a summary of the significance of the global model. If the p-value for the global model is smaller than 0.05, it means that the model explains more of the variations of the response variable than could be expected from random phenomena. In other words, the model is significant at the 5% level. The smaller the p-value, the more significant (and useful) the model.

Second Section: Variables

The second part of the ANOVA table deals with each individual effect (main effects, optionally also interactions and square terms). If the p-value for an effect is smaller than 0.05, it means that the corresponding source of variation explains more of the variations of the response variable than could be expected from random phenomena. In other words, the effect is significant at the 5% level. The smaller the p-value, the more significant the effect.

Model Check

The model check tests whether the non-linear part of the model is significant. It includes up to three groups of effects:

- Interactions (and how they improve a purely linear model);
- Squares (and how they improve a model which already contains interactions);
- Squares (and how they improve a purely linear model).

If the p-value for a group of effects is larger than 0.05, it means that these effects are not useful, and that a simpler model would perform as well. Try to re-compute the response surface without those effects!

Lack of Fit

The lack of fit part tests whether the error in response prediction is mostly due to experimental variability or to an inadequate shape of the model. If the p-value for lack of fit is smaller than 0.05, it means that the model does not describe the true shape of the response surface. In such cases, you may try a transformation of the response variable.

Note that:

1. For ***screening designs***, all terms in the ANOVA table will be ***missing*** if there are as many terms in the model as cube samples (i.e. you have a ***saturated model***). In such cases, you cannot use HOIE for significance testing; try Center samples, Reference samples or COSCIND!
2. If your design has design variables with ***more than two levels***, use ***Multiple Comparisons*** in order to see which levels of a given variable differ significantly from each other.

3. ***Lack of fit*** can only be tested if the ***replicated center samples*** do not all have the same response values (which may sometimes happen by accident).

Classification Table (Table Plot)

This plot shows the classification of each sample. Classes which are significant for a sample are marked with a star (or an asterix).

The outcome of the classification depends on the significance limit; by default it is set to 5%, but you can tune it up or down with the tool.

Look for samples that are not recognized by any of the classes, or those which are allocated to more than one class.

Detailed Effects (Table Plot)

This table gives the numerical values of all effects and their corresponding f-ratios and p-values, for the current response variable. The multiple correlation coefficient and the R-square, which measure the degree of fit of the model, are also presented above the table. A value close to 1 indicates a model with good fit and a value close to 0 indicates bad fit.

Choice of Significance Testing Method

Make sure that you are interpreting the significance of your effects with a relevant significance testing method. Out of the 5 possible methods: HOIE, Center, Reference, Center+Ref, COSCIND, usually only a few are available. Choose HOIE if you have more degrees of freedom in the cube samples than in the Center and/or Reference samples. Choose Center if you want to check the curvature of your response.

Interpreting Effects

This table is particularly useful to display the significance of the effects together with the confounding pattern, for fractional factorial designs where significant effects should be interpreted with caution. If there is any significant effect in your model (p-value smaller than 0.05), check whether this effect has any confounding. If so, you may try an educated guess to find out which of the confounded terms is responsible for the observed effect.

Curvature Check

If you have included replicated center samples in your design, and if you are interpreting your effects with the Center significance testing method, you will also find the p-value for the curvature test above the table. A p-value smaller than 0.05 means that you have a significant curvature: you will need an optimization stage to describe the relationship between your design variables and your response properly.

Effects Overview (Table Plot)

This table plot gives an overview of the significance of all effects for all responses. The sign and significance level of each effect is given as a code:

Significance levels and associated codes

P-value	Negative effect	Positive effect
≥ 0.05	NS	NS
[0.01;0.05]	-	+
[0.005;0.01]	- -	+ +
<0.005	- - -	+ + +

Note: If some of your design variables have more than 2 levels, the Effects Overview table contains stars (*) instead of "+" and "-" signs.

Interpretation: Response Variables

Look for responses which are not significantly explained by any of the design variables: either there are errors in the data, or these responses have very little variation, or they are very noisy, or their variations are caused by non-controlled conditions which have not been included into the design.

Interpretation: Design Variables

Look for rows which contain many "+" or "-" signs: these main effects or interactions dominate. This is how you can detect the most important variables.

Prediction Table (Table Plot)

This table plot shows the predicted values, their deviation, and the reference value (if you predicted with a reference).

You are looking for predictions with as small a deviation as possible. Predictions with high deviations may be outliers.

Predicted vs. Measured (Table Plot)

This table shows the measured and predicted Y values from the response surface model, plus their corresponding X-values and standard error of prediction.

Cross-Correlation (Table Plot)

This table shows the cross-correlations between all variables included in a Statistics analysis.

The table is symmetrical (the correlation between A and B is the same as between B and A) and its diagonal contains only values of 1, since the correlation between a variable and itself is 1.

All other values are between -1 and +1. A large positive value indicates that the corresponding two variables have a tendency to increase simultaneously. A large negative value indicates that when the first variable increases, the other often decreases. A correlation close to 0 indicates that the two variables vary independently from each other.

Special Plots

Interaction Effects (Special Plot)

This plot visualizes the interaction between two design variables.

The plot shows the average response value at the Low and High levels of the first design variable, in two curves: one for the Low level of the second design variable, the other for its High level.

You can see the magnitude of the interaction effect (1/2 * change in the effect of the first design variable when the second design variable changes from Low to High).

- For a ***positive*** interaction, the slope of the effect for "High" is ***larger*** than for "Low";
- For a ***negative*** interaction, the slope of the effect for "High" is ***smaller*** than for "Low".

In addition, the plot also contains information about the value of the interaction effect and its significance (p-value, computed with the significance testing method you have chosen).

Main Effects (Special Plot)

This plot visualizes the main effect of a design variable on a given response.

The plot shows the average response value at the Low and High levels of the design variable. If you have included center samples, the average response value for the center samples is also displayed.

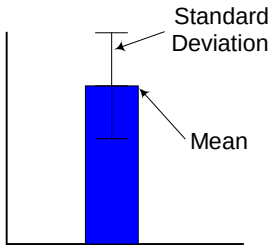
You can see the magnitude of the main effect (change in the response value when the design variable increases from Low to High). If you have center samples, you can also detect a curvature visually.

In addition, the plot also contains information about the value of the effect and its significance (p-value, computed with the significance testing method you have chosen).

Mean and Standard Deviation (Special Plot)

This plot displays the average value and the standard deviation together. The vertical bar is the average value, and the standard deviation is shown as an error bar around the average (see the figure below).

Mean and Sdev for one variable, one group of samples



Interpretation: General Case

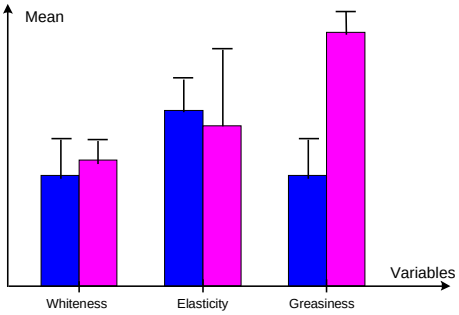
The average response value indicates around which level the values for the various samples are distributed.

The standard deviation is a measure of the spread of the variable around that average. If you are studying several variables together, compare their standard deviations. If standard deviation varies a lot from one variable to another, it will be recommended to standardize the variables in later multivariate analyses (PCA, PLS…). This applies to all kinds of variables except for spectra.

Interpretation: Designed Data

If you have replicated Center samples (or Reference samples), study the Mean and Sdev plot for 2 groups of samples: Design, Center. This enables you to compare the spread over several different experiments (e.g. 16 Design samples) to the spread over a few similar experiments (e.g. 3 Center samples). The former is expected to be much larger than the latter. In the figure below, variables Whiteness and Greasiness have larger spread for the Design samples than the Center samples, which is fine. Variable Elasticity, on the other hand, has a larger spread for its Center samples. This is suspicious: something is probably wrong for one of the Center samples.

Mean and Sdev for 3 responses, with groups "Design samples" and "Center samples"



Multiple Comparisons (Special Plot)

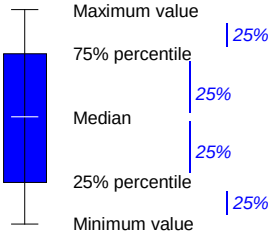
This is a comparison of the average response values for the different levels of a design variable. It tells you which levels of this variable are responsible for a significant change in the response. This plot displays one design variable and one response variable at a time. Look at the plot ID to check which variables are plotted.

- The ***average response value*** is displayed on the left (vertical) axis.
- The names of the different ***levels*** are displayed to the right of the plot, at the same height as the average response value. If a ***reference value*** has been defined in the dialog, it is indicated by ***circles*** to the right of the plot.
- Levels which cannot be distinguished statistically are displayed as points linked by a gray vertical bar. Two levels have ***significantly different*** average response values if they are ***not linked*** by any bar.

Percentiles (Special Plot)

This plot contains one Box-plot for each variable, either over the whole sample set, or for different subgroups. It shows the minimum, the 25% percentile (lower quartile), the median, the 75% percentile (upper quartile) and the maximum.

The box-plot shows 5 percentiles



Note that, if there are less than five samples in the data set, the percentiles are not calculated. The plot then displays one small horizontal bar for each value (each sample). Otherwise, individual samples do not appear on the plot, except for the maximum and minimum values.

Interpretation: General Case

This plot is a good summary of the distributions of your variables. It shows you the total range of variation of each variable. Check whether all variables are within the expected range. If not, out-of-range values are either outliers or data transcription errors. Check your data and correct the errors!

If you have plotted groups of samples (e.g. Design samples, Center samples), there is one box-plot per group.

Check that the **spread** (distance between Min and Max) over the ***Center samples*** is much smaller than the spread over the ***Design samples***. If not, either

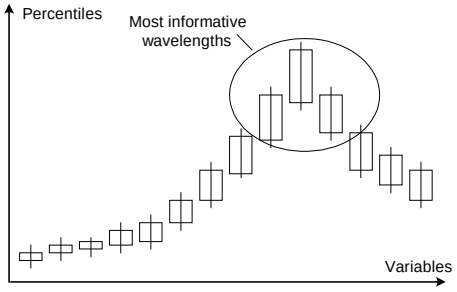
- you have a problem with some of your center samples, or
- this variable has huge uncontrolled variations, or
- this variable has small meaningful variations.

Interpretation: Spectra

This plot can also be used as a diagnostic tool to study the distribution of a whole set of related variables, like in spectroscopy the absorbances for several wavelengths. In such cases, we would recommend not to use subgroups, since otherwise the plot would be too complex to provide interpretable information.

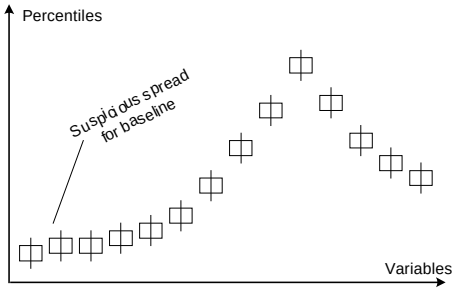
In the figure below, the percentile plot enables you to study the general shape of the spectrum, which is common to all samples in the data set, and also to detect which wavelengths have the largest variation; these are probably the most informative wavelengths.

Percentile plot for variables building up a spectrum



Sometimes, some of the variation may not be relevant to your problem. This is the case in the figure below, which shows an almost uniform spread over all wavelengths. This is very suspicious, since even wavelengths with absorbances close to zero (i.e. baseline) have a large variation over the collected samples. This may indicate a baseline shift, which you can correct using multiplicative scatter correction (MSC). Try to plot scatter effects to check that hypothesis!

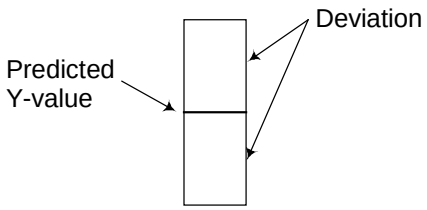
As much variation for the baseline as for the peaks is suspicious



Predicted with Deviations (Special Plot)

This is a plot of predicted Y-value for all prediction samples. The predicted value is shown as a horizontal line. Boxes around the predicted value indicate the deviation, i.e. whether the prediction is reliable or not.

Predicted value and deviation



The deviations are computed as a function of the global model error, the sample leverage, and the sample residual X-variance. A large deviation indicates that the sample used for prediction is not similar to the samples used to make the calibration model. This is a prediction outlier: check its values for the X-variables. If there has been an error, correct it; if the values are correct, the conclusion is that the prediction sample does not belong to the same population as the samples your model is based upon, and you cannot trust the predicted Y value.

Glossary of Terms

2-D Data

This is the most usual data structure in The Unscrambler, as opposed to 3-D data.

3-D Data

Data structure specific to The Unscrambler which accommodates three-way arrays. A 3-D data table can be created from scratch or imported from an external source, then freely manipulated and re-formatted. Note that analyses meant for two-way data structures cannot be run directly on a 3-D data table. You can analyze 3-D X-data together with 2-D Y-data in a Three-Way PLS regression model. If you want to analyze your 3-D data with a 2-way method, duplicate it to a 2-D data layout first.

3-Way PLS

See Three-Way PLS Regression.

Accuracy

The accuracy of a measurement method is its faithfulness, i.e. how close the measured value is to the actual value.

Accuracy differs from precision, which has to do with the spread of successive measurements performed on the same object.

Additive Noise

Noise on a variable is said to be additive when its size is independent of the level of the data value. The range of additive noise is the same for small data values as for larger data values.

Alternating Least Squares

MCR-ALS

Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is an iterative approach (algorithm) to finding the matrices of concentration profiles and pure component spectra from a data table **X** containing the spectra (or instrumental measurements) of several unknown mixtures of a few pure components.

The number of compounds in **X** can be determined using PCA or can be known beforehand. In Multivariate Curve Resolution, it is standard practice to apply MCR-ALS to the same data with varying numbers of components (2 or more).

The MCR-ALS algorithm is described in detail in the Method Reference chapter, available as a separate .PDF document for easy print-out of the algorithms and formulas – download it from Camo's web site www.camo.com/TheUnscrambler/Appendices.

Analysis Of Effects

Calculation of the effects of design variables on the responses. It consists mainly of Analysis of Variance (ANOVA), various Significance Tests, and Multiple Comparisons whenever they apply.

Analysis Of Variance (ANOVA)

Classical method to assess the significance of effects by decomposition of a response's variance into explained parts, related to variations in the predictors, and a residual part which summarizes the experimental error.

The main ANOVA results are: Sum of Squares (SS), number of Degrees of Freedom (DF), Mean Square (MS=SS/DF), F-value, p-value.

The effect of a design variable on a response is regarded as significant if the variations in the response value due to variations in the design variable are large compared with the experimental error. The significance of the effect is given as a p-value: usually, the effect is considered significant if the p-value is smaller than 0.05.

ANOVA

see Analysis of Variance.

Axial Design

One of the three types of mixture designs with a simplex-shaped experimental region. An axial design consists of extreme vertices, overall center, axial points, end points. It can only be used for linear modeling, and therefore it is not available for optimization purposes.

Axial Point

In an axial design, an axial point is positioned on the axis of one of the mixture variables, and must be above the overall center, opposite the end point.

B-Coefficient

See Regression Coefficient.

Bias

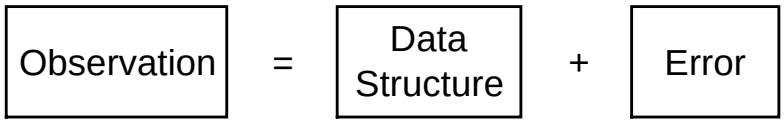
Systematic difference between predicted and measured values. The bias is computed as the average value of the residuals.

Bilinear Modeling

Bilinear modeling (BLM) is one of several possible approaches for data compression.

The bilinear modeling methods are designed for situations where collinearity exists among the original variables. Common information in the original variables is used to build new variables, that reflect the underlying ("latent") structure. These variables are therefore called latent variables. The latent variables are estimated as linear functions of both the original variables and the observations, thereby the name bilinear.

PCA, PCR and PLS are bilinear methods.



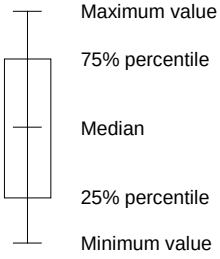
Box-Behnken Design

A class of experimental designs for response surface modeling and optimization, based on only 3 levels of each design variable. The mid-levels of some variables are combined with extreme levels of others. The combinations of only extreme levels (i.e. cube samples of a factorial design) are not included in the design.

Box-Behnken designs are always rotatable. On the other hand, they cannot be built as an extension of an existing factorial design, so they are more recommended when changing the ranges of variation for some of the design variables after a screening stage, or when it is necessary to avoid too extreme situations.

Box-plot

The Box-plot represents the distribution of a variable in terms of percentiles.



Calibration

Stage of data analysis where a model is fitted to the available data, so that it describes the data as good as possible.

After calibration, the variation in the data can be expressed as the sum of a modeled part (structure) and a residual part (noise).

Calibration Samples

Samples on which the calibration is based. The variation observed in the variables measured on the calibration samples provides the information that is used to build the model.

If the purpose of the calibration is to build a model that will later be applied on new samples for prediction, it is important to collect calibration samples that span the variations expected in the future prediction samples.

Category Variable

A category variable is a class variable, i.e. each of its levels is a category (or class, or type), without any possible quantitative equivalent.

Examples: type of catalyst, choice among several instruments, wheat variety, etc..

Candidate Point

In the D-optimal design generation, a number of candidate points are first calculated. These candidate points consist of extreme vertices and centroid points. Then, a number of candidate points is selected D-optimally to create the set of design points.

Center Sample

Sample for which the value of every design variable is set at its mid-level (halfway between low and high).

Center samples have a double purpose: introducing one center sample in a screening design enables curvature checking, and replicating the center sample provides a direct estimation of the experimental error.

Center samples can be included when all design variables are continuous.

Centering

See Mean Centering.

Central Composite Design

A class of experimental designs for response surface modeling and optimization, based on a two-level factorial design on continuous design variables. Star samples and center samples are added to the factorial design, to provide the intermediate levels necessary for fitting a quadratic model.

Central Composite designs have the advantage that they can be built as an extension of a previous factorial design, if there is no reason to change the ranges of variation of the design variables.

If the default star point distance to center is selected, these designs are rotatable.

Centroid Design

See Simplex-centroid design.

Centroid Point

A centroid point is calculated as the mean of the extreme vertices on the design region surface associated with this centroid point. It is used in Simplex-centroid designs, axial designs and D-optimal mixture/non-mixture designs.

Classification

Data analysis method used for predicting class membership. Classification can be seen as a predictive method where the response is a category variable. The purpose of the analysis is to be able to predict which category a new sample belongs to. The main classification method implemented in The Unscrambler is SIMCA classification.

Classification can for instance be used to determine the geographical origin of a raw material from the levels of various impurities, or to accept or reject a product depending on its quality.

To run a classification, you need

- one or several PCA models (one for each class) based on the same variables;
- values of those variables collected on known or unknown samples.

Each new sample is projected onto each PCA model. According to the outcome of this projection, the sample is either recognized as a member of the corresponding class, or rejected.

Closure

In MCR, the Closure constraint forces the sum of the concentrations of all the mixture components to be equal to a constant value (the total concentration) across all samples.

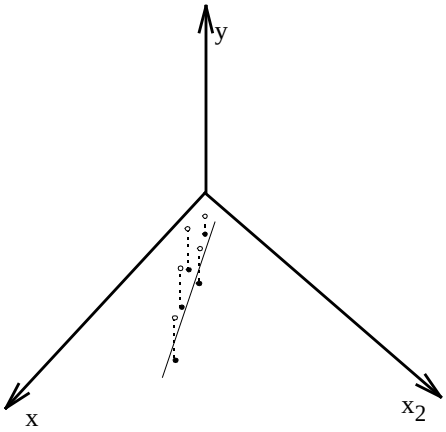
Collinear

See Collinearity.

Collinearity

Linear relationship between variables. Two variables are collinear if the value of one variable can be computed from the other, using a linear relation. Three or more variables are collinear if one of them can be expressed as a linear function of the others.

Variables which are not collinear are said to be linearly independent. Collinearity - or near-collinearity, i.e. very strong correlation - is the major cause of trouble for MLR models, whereas projection methods like PCA, PCR and PLS handle collinearity well.



Component

1) Context: PCA, PCR, PLS… See Principal Component.

2) Context: Curve Resolution: See Pure Components.

3) Context: Mixture Designs: See Mixture Components.

Condition Number

It is the square root of the ratio of the highest eigenvalue to the smallest eigenvalue of the experimental matrix. The higher the condition number, the more spread the region. On the contrary, the lower the condition number, the more spherical the region. The ideal condition number is 1; the closer to 1 the better.

Confounded Effects

Two (or more) effects are said to be confounded when variation in the responses cannot be traced back to the variation in the design variables to which those effects are associated.

Confounded effects can be separated by performing a few new experiments. This is useful when some of the confounded effects have been found significant.

Confounding Pattern

The confounding pattern of an experimental design is the list of the effects that can be studied with this design, with confounded effects listed on the same line.

Constrained Design

Experimental design involving multi-linear constraints between some of the designed variables. There are two types of constrained designed: classical Mixture designs and D-optimal designs.

Constrained Experimental Region

Experimental region which is not only delimited by the ranges of the designed variables, but also by multi-linear constraints existing between these variables. For classical Mixture designs, the constrained experimental region has the shape of a simplex.

Constraint

1) Context: Curve Resolution:

A constraint is a restriction imposed on the solutions to the multivariate curve resolution problem.

Many constraints take the form of a linear relationship between two variables or more:

$$a_1 . X_1 + a_2 . X_2 + ... + a_n . X_n + a_0 >= 0$$

or

$$a_1 . X_1 + a_2 . X_2 + ... + a_n . X_n + a_0 <= 0$$

where X_i are relevant variables (e.g. estimated concentrations), and each constraint is specified by the set of constants $a_0 \ldots a_n$.

2) Context: Mixture Designs: See Multi-Linear Constraint.

Continuous Variable

Quantitative variable measured on a continuous scale.

Examples of continuous variables are:

- Amounts of ingredients (in kg, liters, etc.);

- Recorded or controlled values of process parameters (pressure, temperature, etc.).

Corner Sample

See vertex sample.

Correlation

A unitless measure of the amount of linear relationship between two variables.

The correlation is computed as the covariance between the two variables divided by the square root of the product of their variances. It varies from -1 to +1.

Positive correlation indicates a positive link between the two variables, i.e. when one increases, the other has a tendency to increase too. The closer to +1, the stronger this link.

Negative correlation indicates a negative link between the two variables, i.e. when one increases, the other has a tendency to decrease. The closer to -1, the stronger this link.

Correlation Loadings

Loading plot marking the 50% and 100% explained variance limits. Correlation Loadings are helpful in revealing variable correlations.

COSCIND

A method used to check the significance of effects using a scale-independent distribution as comparison. This method is useful when there are no residual degrees of freedom.

Covariance

A measure of the linear relationship between two variables.

The covariance is given on a scale which is a function of the scales of the two variables, and may not be easy to interpret. Therefore, it is usually simpler to study the correlation instead.

Cross Terms

See Interaction Effects.

Cross Validation

Validation method where some samples are kept out of the calibration and used for prediction. This is repeated until all samples have been kept out once. Validation residual variance can then be computed from the prediction residuals.

In segmented cross validation, the samples are divided into subgroups or “segments”. One segment at a time is kept out of the calibration. There are as many calibration rounds as segments, so that predictions can be made on all samples. A final calibration is then performed with all samples.

In full cross validation, only one sample at a time is kept out of the calibration.

Cube Sample

Any sample which is a combination of high and low levels of the design variables, in experimental plans based on two levels of each variable.

In Box-Behnken designs, all samples which are a combination of high or low levels of some design variables, and center level of others, are also referred to as cube samples.

Curvature

Curvature means that the true relationship between response variations and predictor variations is non-linear.

In screening designs, curvature can be detected by introducing a center sample.

Data Compression

Concentration of the information carried by several variables onto a few underlying variables.

The basic idea behind data compression is that observed variables often contain common information, and that this information can be expressed by a smaller number of variables than originally observed.

Degree Of Fractionality

The degree of fractionality of a factorial design expresses how much the design has been reduced compared to a full factorial design with the same number of variables. It can be interpreted as the number of design variables that should be dropped to compute a full factorial design with the same number of experiments.

Example: with 5 design variables, one can either build

- a full factorial design with 32 experiments (2^5);
- a fractional factorial design with a degree of fractionality of 1, which will include 16 experiments (2^{5-1});
- a fractional factorial design with a degree of fractionality of 2, which will include 8 experiments (2^{5-2}).

Degrees Of Freedom

The number of degrees of freedom of a phenomenon is the number of independent ways this phenomenon can be varied.

Degrees of freedom are used to compute variances and theoretical variable distributions. For instance, an estimated variance is said to be “corrected for degrees of freedom” if it is computed as the sum of square of deviations from the mean, divided by the number of degrees of freedom of this sum.

Design Def Model

In The Unscrambler, predefined set of variables, interactions and squares available for multivariate analyses on Mixture and D-optimal data tables. This set is defined accordingly to the I&S terms included in the model when building the design (**Define Model** dialog).

Design Variable

Experimental factor for which the variations are controlled in an experimental design.

Distribution

Shape of the frequency diagram of a measured variable or calculated parameter. Observed distributions can be represented by a histogram.

Some statistical parameters have a well-known theoretical distribution which can be used for significance testing.

D-Optimal Design

Experimental design generated by the DOPT algorithm. A D-optimal design takes into account the multi-linear relationships existing between design variables, and thus works with constrained experimental regions. There are two types of D-optimal designs: D-optimal Mixture designs and D-optimal Non-Mixture designs, according to the presence or absence of Mixture variables.

D-Optimal Mixture Design

D-optimal design involving three or more Mixture variables and either some Process variables or a mixture region which is not a simplex. In a D-optimal Mixture design, multi-linear relationships can be defined among Mixture variables and/or among Process variables.

D-Optimal Non-Mixture Design

D-optimal design in which some of the Process variables are multi-linearly linked, and which does not involve any Mixture variable.

D-Optimal Principle

Principle consisting in the selection of a sub-set of candidate points which define a maximal volume region in the multi-dimensional space. The D-optimal principle aims at minimizing the condition number.

Edge Center Point

In D-optimal and Mixture designs, the edge center points are positioned in the center of the edges of the experimental region.

End Point

In an axial or a simplex-centroid design, an end point is positioned at the bottom of the axis of one of the mixture variables, and is thus positioned on the side opposite to the axial point.

Experimental Design

Plan for experiments where input variables are varied systematically within predefined ranges, so that their effects on the output variables (responses) can be estimated and checked for significance.

Experimental designs are built with a specific objective in mind, namely screening or optimization.

The number of experiments and the way they are built depends on the objective and on the operational constraints.

Experimental Error

Random variation in the response that occurs naturally when performing experiments.

An estimation of the experimental error is used for significance testing, as a comparison to structured variation that can be accounted for by the studied effects.

Experimental error can be measured by replicating some experiments and computing the standard deviation of the response over the replicates. It can also be estimated as the residual variation when all “structured” effects have been accounted for.

Experimental Region

N-dimensional area investigated in an experimental design with N design variables. The experimental region is defined by:

5. the ranges of variation of the design variables,
7. if any, the multi-linear relationships existing between design variables.

In the case of multi-linear constraints, the experimental region is said to be constrained.

Explained Variance

Share of the total variance which is accounted for by the model.

Explained variance is computed as the complement to residual variance, divided by total variance. It is expressed as a percentage.

For instance, an explained variance of 90% means that 90% of the variation in the data is described by the model, while the remaining 10% are noise (or error).

Explained X-Variance

See Explained Variance.

Explained Y-Variance

See Explained Variance.

F-Distribution

Fisher Distribution is the distribution of the ratio between two variances.

The F-distribution assumes that the individual observations follow an approximate normal distribution.

Fixed Effect

Effect of a variable for which the levels studied in an experimental design are of specific interest.

Examples are:

- effect of the type of catalyst on yield of the reaction;

- effect of resting temperature on bread volume.

The alternative to a fixed effect is a random effect.

Fractional Factorial Design

A reduced experimental plan often used for screening of many variables. It gives as much information as possible about the main effects of the design variables with a minimum of experiments. Some fractional designs also allow two-variable interactions to be studied. This depends on the resolution of the design.

In fractional factorial designs, a subset of a full factorial design is selected so that it is still possible to estimate the desired effects from a limited number of experiments.

The degree of fractionality of a factorial design expresses how fractional it is, compared with the corresponding full factorial.

F-Ratio

The F-ratio is the ratio between explained variance (associated to a given predictor) and residual variance. It shows how large the effect of the predictor is, as compared with random noise.

By comparing the F-ratio with its theoretical distribution (F-distribution), we obtain the significance level (given by a p-value) of the effect.

Full Factorial Design

Experimental design where all levels of all design variables are combined.

Such designs are often used for extensive study of the effects of few variables, especially if some variables have more than two levels. They are also appropriate as advanced screening designs, to study both main effects and interactions, especially if no Resolution V design is available.

Gap

One of the parameters of the Gap-Segment and Norris Gap derivatives, the gap is the length of the interval that separates the two segments that are being averaged.

Look up Segment for more information.

Higher Order Interaction Effects

HOIE is a method to check the significance of effects by using higher order interactions as comparison. This requires that these interaction effects are assumed to be negligible, so that variation associated with those effects is used as an estimate of experimental error.

Histogram

A plot showing the observed distribution of data points. The data range is divided into a number of bins (i.e. intervals) and the number of data points that fall into each bin is summed up.

The height of the bar in the histograms shows how many data points fall within the data range of the bin.

Hotelling T^2 Ellipse

This 95% confidence ellipse can be included in Score plots and reveals potential outliers, lying outside the ellipse. The Hotelling statistic is presented in the Method References chapter, which is available as a .PDF file from CAMO's web site www.camo.com/TheUnscrambler/Appendices .

Influence

A measure of how much impact a single data point (or a single variable) has on the model. The influence depends on the leverage and the residuals.

Inner Relation

In PLS regression models, scores in **X** are used to predict the scores in **Y** and from these predictions, the estimated $\hat{\mathbf{Y}}$ is found. This connection between **X** and **Y** through their scores is called the inner relation.

Interaction

There is an interaction between two design variables when the effect of the first variable depends on the level of the other. This means that the combined effect of the two variables is not equal to the sum of their main effects.

An interaction that increases the main effects is a synergy. If it goes in the opposite direction, it can be called an antagonism.

Intercept

(Also called Offset). The point where a regression line crosses the ordinate (Y-axis).

Interior Point

Point which is not located on the surface, but inside of the experimental region. For example, an axial point is a particular kind of interior point. Interior points are used in classical mixture designs.

Lack Of Fit

In Response Surface Analysis, the ANOVA table includes a special chapter which checks whether the regression model describes the true shape of the response surface. Lack of fit means that the true shape is likely to be different from the shape indicated by the model.

If there is a significant lack of fit, you can investigate the residuals and try a transformation.

Lattice Degree

The degree of a Simplex-Lattice design corresponds to the maximal number of experimental points -1 for a level 0 of one of the Mixture variables.

Lattice Design

See Simplex-lattice design.

Least Square Criterion

Basis of classical regression methods, that consists in minimizing the sum of squares of the residuals. It is equivalent to minimizing the average squared distance between the original response values and the fitted values.

Leveled Variable

A leveled variable is a variable which consists of discrete values instead of a range of continuous values.

Examples are design variables and category variables.

Leveled variables can be used to separate a data table into different groups. This feature is used by the Statistics task, and in sample plots from PCA, PCR, PLS, MLR, Prediction and Classification results.

Levels

Possible values of a variable. A category variable has several levels, which are all possible categories. A design variable has at least a low and a high level, which are the lower and higher bounds of its range of variation. Sometimes, intermediate levels are also included in the design.

Leverage Correction

A quick method to simulate model validation without performing any actual predictions.

It is based on the assumption that samples with a higher leverage will be more difficult to predict accurately than more central samples. Thus a validation residual variance is computed from the calibration sample residuals, using a correction factor which increases with the sample leverage.

Note! For MLR, leverage correction is strictly equivalent to full cross-validation. For other methods, leverage correction should only be used as a quick-and-dirty method for a first calibration, and a proper validation method should be employed later on to estimate the optimal number of components correctly.

Leverage

A measure of how extreme a data point or a variable is compared to the majority.

In PCA, PCR and PLS, leverage can be interpreted as the distance between a projected point (or projected variable) and the model center. In MLR, it is the object distance to the model center.

Average data points have a low leverage. Points or variables with a high leverage are likely to have a high influence on the model.

Limits For Outlier Warnings

Leverage and Outlier limits are the threshold values set for automatic outlier detection. Samples or variables that give results higher than the limits are reported as suspect in the list of outlier warnings.

Linear Effect

See Main Effect.

Linear Model

Regression model including as X-variables the linear effects of each predictor. The linear effects are also called main effects.

Linear models are used in Analysis of Effects in Plackett-Burman and Resolution III fractional factorial designs. Higher resolution designs allow the estimation of interactions in addition to the linear effects.

Loading Weights

Loading weights are estimated in PLS regression. Each X-variable has a loading weight along each model component.

The loading weights show how much each predictor (or X-variable) contributes to explaining the response variation along each model component. They can be used, together with the Y-loadings, to represent the relationship between X- and Y-variables as projected onto one, two or three components (line plot, 2D scatter plot and 3D scatter plot respectively).

Loadings

Loadings are estimated in bilinear modeling methods where information carried by several variables is concentrated onto a few components. Each variable has a loading along each model component.

The loadings show how well a variable is taken into account by the model components. You can use them to understand how much each variable contributes to the meaningful variation in the data, and to interpret variable relationships. They are also useful to interpret the meaning of each model component.

Lower Quartile

The lower quartile of an observed distribution is the variable value that splits the observations into 25% lower values, and 75% higher values. It can also be called 25% percentile.

Main Effect

Average variation observed in a response when a design variable goes from its low to its high level.

The main effect of a design variable can be interpreted as linear variation generated in the response, when this design variable varies and the other design variables have their average values.

MCR

See Multivariate Curve Resolution.

Mean

Average value of a variable over a specific sample set. The mean is computed as the sum of the variable values, divided by the number of samples.

The mean gives a value around which all values in the sample set are distributed. In Statistics results, the mean can be displayed together with the standard deviation.

Mean Centering

Subtracting the mean (average value) from a variable, for each data point.

Median

The median of an observed distribution is the variable value that splits the distribution in its middle: half the observations have a lower value than the median, and the other half have a higher value. It can also be called 50% percentile.

MixSum

Term used in The Unscrambler for ”mixture sum”. See Mixture Sum.

Mixture Components

Ingredients of a mixture.

There must be at least three components to define a mixture. A unique component cannot be called mixture.

Two components mixed together do not require a Mixture design to be studied: study the variation in quantity of one of them as a classical process variable.

Mixture Constraint

Multi-linear constraint between Mixture variables. The general equation for the Mixture constraint is

$$X_1 + X_2 + \ldots + X_n = S$$

where the X_i represent the ingredients of the mixture, and S is the total amount of mixture. In most cases, S is equal to 100%.

Mixture Design

Special type of experimental design, applying to the case of a Mixture constraint. There are three types of classical Mixture designs: Simplex-Lattice design, Simplex-Centroid design, and Axial design. Mixture designs that do not have a simplex experimental region are generated D-optimally; they are called D-optimal Mixture designs.

Mixture Region

Experimental region for a Mixture design. The Mixture region for a classical Mixture design is a simplex.

Mixture Sum

Total proportion of a mixture which varies in a Mixture design. Generally, the mixture sum is equal to 100%. However, it can be lower than 100% if the quantity in one of the components has a fixed value.

The mixture sum can also be expressed as fractions, with values varying from 0 to 1.

Mixture Variable

Experimental factor for which the variations are controlled in a mixture design or D-optimal mixture design. Mixture variables are multi-linearly linked by a special constraint called mixture constraint.

There must be at least three mixture variables to define a mixture design. See Mixture Components.

MLR

See Multiple Linear Regression.

Mode

See Modes.

Model

Mathematical equation summarizing variations in a data set.

Models are built so that the structure of a data table can be understood better than by just looking at all raw values.

Statistical models consist of a structure part and an error part. The structure part (information) is intended to be used for interpretation or prediction, and the error part (noise) should be as small as possible for the model to be reliable.

Model Center

The model center is the origin around which variations in the data are modeled. It is the (0,0) point on a score plot.

If the variables have been centered, samples close to the average will lie close to the model center.

Model Check

In Response Surface Analysis, a section of the ANOVA table checks how useful the interactions and squares are, compared with a purely linear model. This section is called Model Check.

If one part of the model is not significant, it can be removed so that the remaining effects are estimated with a better precision.

Modes

In a multi-way array, a mode is one of the structuring dimensions of the array. A two-way array (standard n x p matrix) has two modes: rows and columns. A three-way array (3-D data table, or some result matrices) has three modes: rows, columns and planes – or e.g. Samples, Primary variables and Secondary variables.

Multiple Comparison Tests

Tests showing which levels of a category design variables can be regarded as causing real differences in response values, compared to other levels of the same design variable.

For continuous or binary design variables, analysis of variance is sufficient to detect a significant effect and interpret it. For category variables, a problem arises from the fact that, even when analysis of variance shows a significant effect, it is impossible to know which levels are significantly different from others. This is why multiple comparisons have been implemented. They are to be used once analysis of variance has shown a significant effect for a category variable.

Multi-Linear Constraint

This is a linear relationship between two variables or more. A constraint has the general form:

$$A_1 . X_1 + A_2 . X_2 + ... + A_n . X_n + A_0 >= 0$$

or

$$A_1 . X_1 + A_2 . X_2 + ... + A_n . X_n + A_0 <= 0$$

where X_i are designed variables (mixture or process), and each constraint is specified by the set of constants $A_0 \ldots A_n$.

A multi-linear constraint cannot involve both Mixture and Process variables.

Multi-Way Analysis

See Three-Way PLS Regression.

Multi-Way Data

See 3-D Data.

Multiple Linear Regression (MLR)

A method for relating the variations in a response variable (Y-variable) to the variations of several predictors (X-variables), with explanatory or predictive purposes.

An important assumption for the method is that the X-variables are linearly independent, i.e. that no linear relationship exists between the X-variables. When the X-variables carry common information, problems can arise due to exact or approximate collinearity.

Multivariate Curve Resolution (MCR)

A method that resolves unknown mixtures into n pure components. The number of components and their concentrations and instrumental profiles are estimated in a way that explains the structure of the observed data under the chosen model constraints.

Noise

Random variation that does not contain any information.

The purpose of multivariate modeling is to separate information from noise.

Non-Linearity

Deviation from linearity in the relationship between a response and its predictors.

Non-Negativity

In MCR, the Non-negativity constraint forces the values in a profile to be equal to or greater than zero.

Normal Distribution

Frequency diagram showing how independent observations, measured on a continuous scale, would be distributed if there were an infinite number of observations and no factors caused systematic effects.

A normal distribution can be described by two parameters:

- a theoretical mean, which is the center of the distribution;
- a theoretical standard deviation, which is the spread of the individual observations around the mean.

Normal Probability Plot

The normal probability plot (or N-plot) is a 2-D plot which displays a series of observed or computed values in such a way that their distribution can be visually compared to a normal distribution.

The observed values are used as abscissa, and the ordinate displays the corresponding percentiles on a special scale. Thus if the values are approximately normally distributed around zero, the points will appear close to a straight line going through (0,50%).

A normal probability plot can be used to check the normality of the residuals (they should be normal; outliers will stick out), and to visually detect significant effects in screening designs with few residual degrees of freedom.

NPLS

See Three-Way PLS Regression.

O^2V

In The Unscrambler, three-way data structure formed of two Object modes and one Variable mode. A 3-D data table with layout O^2V is displayed in the Editor as a "flat" (unfolded) table with as many rows as Primary samples times Secondary samples and as many columns as Variables.

Offset

See Intercept.

Optimization

Finding the settings of design variables that generate optimal response values.

Orthogonal

Two variables are said to be orthogonal if they are completely uncorrelated, i.e. their correlation is 0.

In PCA and PCR, the principal components are orthogonal to each other.

Factorial designs, Plackett-Burman designs, Central Composite designs and Box-Behnken designs are built in such a way that the studied effects are orthogonal to each other.

Orthogonal Design

Designs built in such a way that the studied effects are orthogonal to each other, are called orthogonal designs.

Examples: Factorial designs, Plackett-Burman designs, Central Composite designs and Box-behnken designs.

D-optimal designs and classical mixture designs are not orthogonal.

Outlier

An observation (outlying sample) or variable (outlying variable) which is abnormal compared to the major part of the data.

Extreme points are not necessarily outliers; outliers are points that apparently do not belong to the same population as the others, or that are badly described by a model.

Outliers should be investigated before they are removed from a model, as an apparent outlier may be due to an error in the data.

OV^2

In The Unscrambler, three-way data structure formed of one Object mode and two Variable modes. A 3-D data table with layout OV^2 is displayed in the Editor as a "flat" (unfolded) table with as many rows as Objects (samples) and as many columns as Primary variables times Secondary variables.

Overfitting

For a model, overfitting is a tendency to describe too much of the variation in the data, so that not only consistent structure is taken into account, but also some noise or uninformative variation.

Overfitting should be avoided, since it usually results in a lower quality of prediction. Validation is an efficient way to avoid model overfitting.

Partial Least Squares Regression

See PLS Regression.

Passified

When you apply the "Passify" weighting option to a variable, it becomes Passified. This means that it loses all influence on the model, but it is not removed from the analysis, so that you can study how it correlates to the other variables, by plotting Correlation Loadings.

Variables which are not passified may be called "active variables".

Passify

New weighting option which allows you, by giving a variable a very low weight in a PCA, PCR or PLS model, to remove its influence on the model while still showing how it correlates to other variables.

PCA

See Principal Component Analysis.

PCR

See Principal Component Regression.

PCs

See Principal Component.

Percentile

The X% percentile of an observed distribution is the variable value that splits the observations into X% lower values, and 100-X% higher values.

Quartiles and median are percentiles. The percentiles are displayed using a box-plot.

Plackett-Burman Design

A very reduced experimental plan used for a first screening of many variables. It gives information about the main effects of the design variables with the smallest possible number of experiments.

No interactions can be studied with a Plackett-Burman design, and moreover, each main effect is confounded with a combination of several interactions, so that these designs should be used only as a first stage, to check whether there is any meaningful variation at all in the investigated phenomena.

PLS

See PLS Regression.

PLS Discriminant Analysis (PLS-DA)

Classification method based on modeling the differences between several classes with PLS.

If there are only two classes to separate, the PLS model uses one response variable, which codes for class membership as follows: -1 for members of one class, +1 for members of the other one. The PLS1 algorithm is then used.

If there are three classes or more, PLS2 is used, with one response variable (-1/+1 or 0/1, which is equivalent) coding for each class.

PLS Regression (PLS)

A method for relating the variations in one or several response variables (Y-variables) to the variations of several predictors (X-variables), with explanatory or predictive purposes.

This method performs particularly well when the various X-variables express common information, i.e. when there is a large amount of correlation, or even collinearity.

Partial Least Squares Regression is a bilinear modeling method where information in the original X-data is projected onto a small number of underlying ("latent") variables called PLS components. The Y-data are actively used in estimating the "latent" variables to ensure that the first components are those that are most relevant for predicting the Y-variables. Interpretation of the relationship between X-data and Y-data is then simplified as this relationship in concentrated on the smallest possible number of components.

By plotting the first PLS components one can view main associations between X-variables and Y-variables, and also interrelationships within X-data and within Y-data.

PLS1

Version of the PLS method with only one Y-variable.

PLS2

Version of the PLS method in which several Y-variables are modeled simultaneously, thus taking advantage of possible correlations or collinearity between Y-variables.

PLS-DA

See PLS Discriminant Analysis.

Precision

The precision of an instrument or a measurement method is its ability to give consistent results over repeated measurements performed on the same object. A precise method will give several values that are very close to each other.

Precision can be measured by standard deviation over repeated measurements.

If precision is poor, it can be improved by systematically repeating the measurements over each sample, and replacing the original values by their average for that sample.

Precision differs from accuracy, which has to do with how close the average measured value is to the target value.

Prediction

Computing response values from predictor values, using a regression model.

To make predictions, you need

- a regression model (PCR or PLS), calibrated on X- and Y-data;
- new X-data collected on samples which should be similar to the ones used for calibration.

The new X-values are fed into the model equation (which uses the regression coefficients), and predicted Y-values are computed.

Predictor

Variable used as input in a regression model. Predictors are usually denoted X-variables.

Primary Sample

In a 3-D data table with layout O^2V, this is the major Sample mode. Secondary samples are nested within each Primary sample.

Primary Variable

In a 3-D data table with layout OV^2, this is the major Variable mode. Secondary variables are nested within each Primary variable.

Principal Component Analysis (PCA)

PCA is a bilinear modeling method which gives an interpretable overview of the main information in a multidimensional data table.

The information carried by the original variables is projected onto a smaller number of underlying ("latent") variables called principal components. The first principal component covers as much of the variation in the data as possible. The second principal component is orthogonal to the first and covers as much of the remaining variation as possible, and so on.

By plotting the principal components, one can view interrelationships between different variables, and detect and interpret sample patterns, groupings, similarities or differences.

Principal Component Regression (PCR)

PCR is a method for relating the variations in a response variable (Y-variable) to the variations of several predictors (X-variables), with explanatory or predictive purposes.

This method performs particularly well when the various X-variables express common information, i.e. when there is a large amount of correlation, or even collinearity.

Principal Component Regression is a two-step method. First, a Principal Component Analysis is carried out on the X-variables. The principal components are then used as predictors in a Multiple Linear Regression.

Principal Component (PC)

Principal Components (PCs) are composite variables, i.e. linear functions of the original variables, estimated to contain, in decreasing order, the main structured information in the data. A PC is the same as a score vector, and is also called a latent variable.

Principal components are estimated in PCA and PCR. PLS components are also denoted PCs.

Process Variable

Experimental factor for which the variations are controlled in an experimental design, and to which the mixture variable definition does not apply.

Projection

Principle underlying bilinear modeling methods such as PCA, PCR and PLS.

In those methods, each sample can be considered as a point in a multi-dimensional space. The model will be built as a series of components onto which the samples - and the variables - can be projected. Sample projections are called scores, variable projections are called loadings.

The model approximation of the data is equivalent to the orthogonal projection of the samples onto the model. The residual variance of each sample is the squared distance to its projection.

Proportional Noise

Noise on a variable is said to be proportional when its size depends on the level of the data value. The range of proportional noise is a percentage of the original data values.

Pure Components

In MCR, an unknown mixture is resolved into **n** pure components. The number of components and their concentrations and instrumental profiles are estimated in a way that explains the structure of the observed data under the chosen model constraints.

p-value

The p-value measures the probability that a parameter estimated from experimental data should be as large as it is, if the real (theoretical, non-observable) value of that parameter were actually zero. Thus, p-value is used to assess the significance of observed effects or variations: a small p-value means that you run little risk of mistakenly concluding that the observed effect is real.

The usual limit used in the interpretation of a p-value is 0.05 (or 5%). If p-value < 0.05, you have reason to believe that the observed effect is not due to random variations, and you may conclude that it is a significant effect.

p-value is also called “significance level”.

Quadratic Model

Regression model including as X-variables the linear effects of each predictor, all two-variable interactions, and the square effects.

With a quadratic model, the curvature of the response surface can be approximated in a satisfactory way.

Random Effect

Effect of a variable for which the levels studied in an experimental design can be considered to be a small selection of a larger (or infinite) number of possibilities.

Examples:

- Effect of using different batches of raw material;

- Effect of having different persons perform the experiments.

The alternative to a random effect is a fixed effect.

Random Order

Randomization is the random mixing of the order in which the experiments are to be performed. The purpose is to avoid systematic errors which could interfere with the interpretation of the effects of the design variables.

Reference Sample

Sample included in a designed data table to compare a new product under development to an existing product of a similar type.

The design file will contain only response values for the reference samples, whereas the input part (the design part) is missing (m).

Regression Coefficient

In a regression model equation, regression coefficients are the numerical coefficients that express the link between variation in the predictors and variation in the response.

Regression

Generic name for all methods relating the variations in one or several response variables (Y-variables) to the variations of several predictors (X-variables), with explanatory or predictive purposes.

Regression can be used to describe and interpret the relationship between the X-variables and the Y-variables, and to predict the Y-values of new samples from the values of the X-variables.

Repeated Measurement

Measurement performed several times on one single experiment or sample.

The purpose of repeated measurements is to estimate the measurement error, and to improve the precision of an instrument or measurement method by averaging over several measurements.

Replicate

Replicates are experiments that are carried out several times. The purpose of including replicates in a data table is to estimate the experimental error.

Replicates should not be confused with repeated measurements, which give information about measurement error.

Residual

A measure of the variation that is not taken into account by the model.

The residual for a given sample and a given variable is computed as the difference between observed value and fitted (or projected, or predicted) value of the variable on the sample.

Residual Variance

The mean square of all residuals, sample- or variable-wise.

This is a measure of the error made when observed values are approximated by fitted values, i.e. when a sample or a variable is replaced by its projection onto the model.

The complement to residual variance is explained variance.

Residual X-Variance

See Residual Variance.

Residual Y-Variance

See Residual Variance.

Resolution

1) Context: experimental design

Information on the degree of confounding in fractional factorial designs.

Resolution is expressed as a roman number, according to the following code:

- in a Resolution III design, main effects are confounded with 2-factor interactions;

- in a Resolution IV design, main effects are free of confounding with 2-factor interactions, but 2-factor interactions are confounded with each other;
- in a Resolution V design, main effects and 2-factor interactions are free of confounding.

More generally, in a Resolution R design, effects of order k are free of confounding with all effects of order less than R-k.

2) Context: data analysis

Extraction of estimated pure component profiles and spectra from a data matrix. See Multivariate Curve Resolution for more details.

Response Surface Analysis

Regression analysis, often performed with a quadratic model, in order to describe the shape of the response surface precisely.

This analysis includes a comprehensive ANOVA table, various diagnostic tools such as residual plots, and two different visualizations of the response surface: contour plot and landscape plot.

Note: Response surface analysis can be run on designed or non-designed data. However it is not available for Mixture Designs; use PLS instead.

Response Variable

Observed or measured parameter which a regression model tries to predict.

Responses are usually denoted Y-variables.

Responses

See Response Variable.

RMSEC

Root Mean Square Error of Calibration. A measurement of the average difference between predicted and measured response values, at the calibration stage.

RMSEC can be interpreted as the average modeling error, expressed in the same units as the original response values.

RMSED

Root Mean Square Error of Deviations. A measurement of the average difference between the abscissa and ordinate values of data points in any 2D scatter plot.

RMSEP

Root Mean Square Error of Prediction. A measurement of the average difference between predicted and measured response values, at the prediction or validation stage.

RMSEP can be interpreted as the average prediction error, expressed in the same units as the original response values.

Sample

Object or individual on which data values are collected, and which builds up a row in a data table.

In experimental design, each separate experiment is a sample.

Scaling

See Weighting.

Scatter Effects

In spectroscopy, scatter effects are effects that are caused by physical phenomena, like particle size, rather than chemical properties. They interfere with the relationship between chemical properties and shape of the spectrum. There can be additive and multiplicative scatter effects.

Additive and multiplicative effects can be removed from the data by different methods. Multiplicative Scatter Correction removes the effects by adjusting the spectra from ranges of wavelengths supposed to carry no specific chemical information.

Scores

Scores are estimated in bilinear modeling methods where information carried by several variables is concentrated onto a few underlying variables. Each sample has a score along each model component.

The scores show the locations of the samples along each model component, and can be used to detect sample patterns, groupings, similarities or differences.

Screening

First stage of an investigation, where information is sought about the effects of many variables. Since many variables have to be investigated, only main effects, and optionally interactions, can be studied at this stage.

There are specific experimental designs for screening, such as factorial or Plackett-Burman designs.

Secondary Sample

In a 3-D data table with layout O^2V, this is the minor Sample mode. Secondary samples are nested within each Primary sample.

Secondary Variable

In a 3-D data table with layout OV^2, this is the minor Variable mode. Secondary variables are nested within each Primary variable.

Segment

One of the parameters of Gap-Segment derivatives and Moving Average smoothing, a segment is an interval over which data values are averaged.

In smoothing, X-values are averaged over ***one segment*** symmetrically surrounding a data point. The raw value on this point is replaced by the average over the segment, thus creating a ***smoothing*** effect.

In Gap-Segment derivatives (designed by Karl Norris), X-values are averaged separately over ***one segment on each side*** of the data point. The two segments are separated by a gap. The raw value on this point is replaced by the difference of the two averages, thus creating an estimate of the ***derivative*** on this point.

Sensitivity to Pure Components

In MCR computations, Sensitivity to Pure Components is one of the parameters influencing the convergence properties of the algorithm. It can be roughly interpreted as how dominating the last estimated primary principal component is (the one that generates the weakest structure in the data), compared to the first one.

The ***higher*** the sensitivity, the ***more*** pure components will be extracted.

SEP

See Standard Error of Performance.

Significance Level

See p-value.

Significant

An observed effect (or variation) is declared significant if there is a small probability that it is due to chance.

SIMCA

See SIMCA Classification.

SIMCA Classification

Classification method based on disjoint PCA modeling.

SIMCA focuses on modeling the similarities between members of the same class. A new sample will be recognized as a member of a class if it is similar enough to the other members; else it will be rejected.

Simplex

Specific shape of the experimental region for a classical mixture design. A Simplex has N corners but N-1 independent variables in a N-dimensional space. This results from the fact that whatever the proportions of the ingredients in the mixture, the total amount of mixture has to remain the same: the N^{th} variable depends on the N-1 other ones. When mixing three components, the resulting simplex is a triangle.

Simplex-Centroid Design

One of the three types of mixture designs with a simplex-shaped experimental region. A Simplex-centroid design consists of extreme vertices, center points of all "sub-simplexes", and the overall center. A "sub-simplex" is a simplex defined by a subset of the design variables. Simplex-centroid designs are available for optimization purposes, but not for a screening of variables.

Simplex-Lattice Design

One of the three types of mixture designs with a simplex-shaped experimental region. A Simplex-lattice design is a mixture variant of the full-factorial design. It is available for both screening and optimization purposes, according to the degree of the design (See lattice degree).

Square Effect

Average variation observed in a response when a design variable goes from its center level to an extreme level (low or high).

The square effect of a design variable can be interpreted as the curvature observed in the response surface, with respect to this particular design variable.

Standard Deviation

Sdev is a measure of a variable's spread around its mean value, expressed in the same unit as the original values.

Standard deviation is computed as the square root of the mean square of deviations from the mean.

Standard Error Of Performance (SEP)

Variation in the precision of predictions over several samples.

SEP is computed as the standard deviation of the residuals.

Standardization

Widely used pre-processing that consists in first centering the variables, then scaling them to unit variance.

The purpose of this transformation is to give all variables included in an analysis an equal chance to influence the model, regardless of their original variances.

In The Unscrambler, standardization can be performed automatically when computing a model, by choosing 1/SDev as variable weights.

Star Points Distance To Center

In Central Composite designs, the properties of the design vary according to the distance between the star samples and the center samples. This distance is measured in normalized units, i.e. assuming that the low cube level of each variable is -1 and the high cube level +1.

Three cases can be considered:

- The default star distance to center ensures that all design samples are located on the surface of a sphere. In other words, the star samples are as far away from the center as the cube samples are. As a consequence, all design samples have exactly the same leverage. The design is said to be "rotatable";
- The star distance to center can be tuned down to 1. In that case, the star samples will be located at the centers of the faces of the cube. This ensures that a Central Composite design can be built even if levels lower than "low cube" or higher than "high cube" are impossible. However, the design is no longer rotatable;
- Any intermediate value for the star distance to center is also possible. The design will not be rotatable.

Star Samples

In optimization designs of the Central Composite family, star samples are samples with mid-values for all design variables except one, for which the value is extreme. They provide the necessary intermediate levels that will allow a quadratic model to be fitted to the data.

Star samples can be centers of cube faces, or they can lie outside the cube, at a given distance (larger than 1) from the center of the cube – see Star Points Distance To Center.

Steepest Ascent

On a regular response surface, the shortest way to the optimum can be found by using the direction of steepest ascent.

Student t-distribution

=t-distribution. Frequency diagram showing how independent observations, measured on a continuous scale, are distributed around their mean when the mean and standard deviation have been estimated from the data and when no factor causes systematic effects.

When the number of observations increases towards an infinite number, the Student t-distribution becomes identical to the normal distribution.

A Student t-distribution can be described by two parameters: the mean value, which is the center of the distribution, and the standard deviation, which is the spread of the individual observations around the mean. Given those two parameters, the shape of the distribution further depends on the number of degrees of freedom, usually n-1, if n is the number of observations.

Test Samples

Additional samples which are not used during the calibration stage, but only to validate an already calibrated model.

The data for those samples consist of X-values (for PCA) or of both X- and Y-values (for regression). The model is used to predict new values for those samples, and the predicted values are then compared to the observed ones.

Test Set Validation

Validation method based on the use of different data sets for calibration and validation. During the calibration stage, calibration samples are used. Then the calibrated model is used on the test samples, and the validation residual variance is computed from their prediction residuals.

Three-Way PLS

See Three-Way PLS Regression.

Three-Way PLS Regression

A method for relating the variations in one or several response variables (Y-variables) arranged in a 2-D table to the variations of several predictors arranged in a 3-D table (Primary and Secondary X-variables), with explanatory or predictive purposes.

See PLS Regression for more details.

Training Samples

See Calibration Samples.

Tri-PLS

See Three-Way PLS Regression.

T-Scores

The scores found by PCA, PCR and PLS in the X-matrix.

See Scores for more details.

Tukey´s Test

A multiple comparison test (see Multiple Comparison Tests for more details).

t-value

The t-value is computed as the ratio between deviation from the mean accounted for by a studied effect, and standard error of the mean.

By comparing the t-value with its theoretical distribution (Student t-distribution), we obtain the significance level of the studied effect.

UDA

See User-Defined Analysis.

UDT

See User-Defined Transformation.

Uncertainty Limits

Limits produced by Uncertainty Testing, helping you assess the significance of your X-variables in a regression model. Variables with uncertainty limits that do not cross the "0" axis are significant.

Uncertainty Test

Martens' Uncertainty Test is a significance testing method implemented in The Unscrambler, which assesses the stability of PCA or Regression results. Many plots and results are associated to the test, allowing the estimation of the model stability, the identification of perturbing samples or variables, and the selection of significant X-variables. The test is performed with Cross Validation, and is based on the Jack-knifing principle.

Underfit

A model that leaves aside some of the structured variation in the data is said to underfit.

Unfold

Operation consisting in mapping a three-way data structure onto a "flat", two-way layout. An unfolded three-way array has one of its original modes nested into another one. In horizontal unfolding, all planes are displayed side by side, resulting in an OV^2 layout, with Primary and Secondary variables. In vertical unfolding, all planes are displayed on top of each other, resulting in an O^2V layout, with Primary and Secondary samples.

Unimodality

In MCR, the Unimodality constraint allows the presence of only one maximum per profile.

Upper Quartile

The upper quartile of an observed distribution is the variable value that splits the observations into 75% lower values, and 25% higher values. It can also be called 75% percentile.

U-Scores

The scores found by PLS in the Y-matrix.

See Scores for more details.

User-Defined Analysis (UDA)

DLL routine programmed in C++, Visual Basic, Matlab or other. UDAs allow the user to program his own analysis methods and use them in The Unscrambler.

User-Defined Transformation (UDT)

DLL routine programmed in C++, Visual Basic, Matlab or other. UDTs allow the user to program his own pre-processing methods and use them in The Unscrambler.

Validation Samples

See Test Samples.

Validation

Validation means checking how well a model will perform for future samples taken from the same population as the calibration samples. In regression, validation also allows for estimation of the prediction error in future predictions.

The outcome of the validation stage is generally expressed by a validation variance. The closer the validation variance is to the calibration variance, the more reliable the model conclusions.

When explained validation variance stops increasing with additional model components, it means that the noise level has been reached. Thus the validation variance is a good diagnostic tool for determining the proper number of components in a model.

Validation variance can also be used as a way to determine how well a single variable is taken into account in an analysis. A variable with a high explained validation variance is reliably modeled and is probably quite precise; a variable with a low explained validation variance is badly taken into account and is probably quite noisy.

Three validation methods are available in The Unscrambler:

- test set validation;
- cross validation;
- leverage correction.

Variable

Any measured or controlled parameter that has varying values over a given set of samples.

A variable determines a column in a data table.

Variance

A measure of a variable's spread around its mean value, expressed in square units as compared to the original values.

Variance is computed as the mean square of deviations from the mean. It is equal to the square of the standard deviation.

Vertex Sample

A vertex is a point where two lines meet to form an angle. Vertex samples are used in Simplex-centroid, axial and D-optimal mixture/non-mixture designs.

Ways

See Modes.

Weighting

A technique to modify the relative influences of the variables on a model. This is achieved by giving each variable a new weight, ie. multiplying the original values by a constant which differs between variables. This is also called scaling.

The most common weighting technique is standardization, where the weight is the standard deviation of the variable.

Index

2

2-D 235
2-D data 235
2D scatter plot 59

3

3-D 235
3-D data 235
  in the Editor 84
  unfold 52
3-D data table
  O2V 52
  OV2 52
  OV2 vs. O2V 52
3-D layout 251, 252, 263
3D scatter plot 59

A

absorbance to reflectance 74
accuracy 235
additive noise 235
alternating least squares 167, 235
analysis
  constrained experiments 152
Analysis
  Constrained Experiments 152
analysis of designed data 147
analysis of effects 148, 236
analysis of variance 148. *See* ANOVA
ANOVA 148, 236, 246, 249
  for linear response surfaces 151
  for quadratic response surfaces 151
  linear 148
  linear with interactions 148
  quadratic 148, 149
  summary 148
  table plot interpretation 227
area normalization 72
averaging 80
axial design 236
axial point 236

B

badly described variables
  X 200
  Y 202
b-coefficient 256
B-coefficient 236, 256
b-coefficients 151
  standard error 151
B-coefficients 109, 151
  standard error 151
bias 236
bi-linear modeling 236
binary variables 17
BLM 236
blocking 44
box plots 90
Box-Behnken design 24, 237
box-plot 232, 237
build a non-designed data table 54
build an experimental design 55

C

calibration 108, 237
calibration samples 237
candidate point 238
category variable 237
category variables 17
  binary variables 17
  levels 17
center sample 238, 241
center samples 23, 40, 149
centering 80, 238
  three-way data 83
central composite design 238
  center samples 23
  cube samples 23
  star samples 23
central composite designs 23
centroid design 238
centroid point 238
classification 135, 238
  Cooman's plot 138
  discriminant analysis 138
  discrimination power 137
  Hi 137
  model distance 137
  modeling power 137
  project onto regression model 138
  scores (plot) 202
  Si 137
  Si vs. Hi 138
  SIMCA 135, 260
  SIMCA modeling 136
  table plot interpretation 228

classification scores
plot interpretation 202
classify
new samples 136
close file 55
closure 239
clustering 14
find groups of samples 212, 221
clustering results 145
collinear 239
collinearity 239
comparison with scale-independent distribution 149.
See COSIND
component 239
condition number 239
confounded effects 239
confounding 20, 21, 257
confounding pattern 20, 22, 240
constrained design 240
constrained experimental region 240
constraint 240
closure 164
cost 50
non-negativity 164
other constraints in MCR 165
unimodality 164
Constraint 240
closure 164
Cost 50
non-negativity 164
other constraints in MCR 164
unimodality 164
constraints
MCR 163
continuous variable 16
continuous variables 240
levels 16, 17
contour plot 151
Cooman's plot 138
interpretation 202
core array 180
corner sample 240
correlation 240
correlation between variables
interpretation 206
interpretation, loading plot 205
correlation loadings 241
interpretation 206, 207, 208
COSCIND 149, 241
covariance 241
create a data table 53
cross terms 241
cross validation 120
full 120, 121
segmented 120, 121
test-set switch 120
cross-correlation
matrix plot interpretation 225
table plot interpretation 230
cross-validation 241
cube sample 39, 241
cube samples 23
curvature 40, 241
check 40
detect 189, 229

D

data compression 241
data tables, create by import 55
data tables, create new 53
data tables, create new designed 55
data tables, create new non-designed 54
degree of fractionality 242
degrees of freedom 148, 242
derivatives 76
gap 245
gap-segment 76
Norris-gap 76
Savitzky-Golay 76
segment 259
descriptive multivariate analysis 93
descriptive statistics 89
2D scatter plots 90
box plots 90
line plots 90
plots 90
descriptive variable analysis 90
design 16
Box-Behnken 24
category variables 17
center samples 40
central composite 23
continuous variables 16
design variables 16
D-optimal mixture 242
D-optimal non-mixture 243
D-Optimal Non-Mixture 243
extend 44
fractional factorial 20, 242, 244
full factorial 19, 244
mixture 248
Mixture 248
mixture variables 17
non-design variables 17
orthogonal 252
Plackett-Burman 22, 253
process variables 18
reference samples 42
replicates 42
resolution 20, 22
screening 19

simplex-centroid 260
simplex-lattice 260
types 18
Design Def model 242
design variable 242
design variables 16, 47
category variables 17
continuous variables 16
select 47
designed data 13
detailed effects
plot interpretation 185
table plot interpretation 229
detect
curvature 189, 229
lack of fit 228
outlier 213, 217, 218, 219, 222
significant effects 228, 229
detect lack
of fit 227
detect non-linearities 113
detect outlier 227
deviations
interpretation 233
df 148. *See* degrees of freedom
differentiation 76
discrimination power 137
plot interpretation 185
distribution 242, 245
normal 251
visualize 61
D-optimal design 242
PLS analysis 152
D-Optimal Design 242
PLS analysis 152
D-optimal mixture design 242
D-optimal non-mixture design 243
D-Optimal Non-Mixture Design 243
D-optimal principle 243
D-Optimal Principle 28, 29, 243

E

edge center point 243
editing operations 69
effects
find important 226
n-plot 226
significance 228, 229
effects overview
plot interpretation 229
EMSC 75
end point 243
error measures 110
estimated concentrations 162
plot interpretation 185
estimated spectra 162
plot interpretation 186
experimental design 243
experimental design, create new 55
experimental error 243
Experimental error 243
experimental region 243
Experimental Region 243
experimental strategy 46
explained variance 95, 98, 243
explained Y-variance 110
extend
designs 44

F

factors 16
F-distribution 244
file properties 55
Fisher distribution 244
fixed effect 244
fractional design
resolution 20, 22
fractional factorial design 20, 240, 242, 244
f-ratio 148, 244
F-ratio 244
f-ratios
plot interpretation 186
full cross validation 120, 121
full factorial design 19, 244

G

gap 245
gap-segment derivatives 76
gaussian filtering 70, 71
group selection of test set 119, 120
groups
find groups of samples 212, 221

H

Hi 137
higher order interaction effects 149, 245
histogram 61, 242, 245
preference ratings 65
results 66
HOIE 149, 245
Hotelling T2 ellipse 245

I

import data 55
influence 245
plot interpretation 203, 204, 211, 220
influential outlier 217, 218, 219

influential samples 204, 205
inner relation 245
  tri-PLS 181
interaction 245
interaction effects
  plot interpretation 230
interactions 18
intercept 245
interior point 246
interpret
  PCA 99

J

jack-knifing 121, 127. *See* uncertainty test

K

Kubelka-Munk 74

L

lack of fit 151, 246
  detect 227, 228
  in regression 113. *See* non-linearities
landscape plot 151
lattice degree 246
lattice design 246
least square criterion 246
least squares 246
leveled variables 246
levels 246
levels of continuous variables 16, 17
leverage 245, 247
  correction 120
leverage correction 246
leverages
  designed data 187, 203, 205
  high-leverage sample 187
  influential samples 204, 205
  interpretation, influence plot 203, 204
  plot interpretation 186, 222
limits for outlier warnings 247
line plot 58, 90
linear effect 247
linear model 247
loading weights 111, 247
  plot interpretation 189, 208, 209, 221
  plot interpretation (tri-PLS) 208, 209
  uncertainty 122
loadings 96, 247
  p-loadings 111
  plot interpretation 187, 188, 205, 206, 207, 220, 221
  PLS 111
  q-loadings 111
  uncertainty 122
logarithmic transformation 70
lower quartile 247

M

main effect 247
main effects 18
  plot interpretation 231
manual selection of test set 119, 120
Martens' Uncertainty Test 121
matrix
  plot 60
matrix plot
  3-D 64
maximizing single responses 19
maximum normalization 73
MCR 248
  algorithm 167, 235
  ambiguity 163
  applications 166
  co-elution 166
  comparison with PCA 160
  constraints 163
  estimated concentrations 162
  estimated spectra 162
  initial guess 167
  non-unique solution 163
  number of components 161
  purposes 160
  residuals 162
  sample residuals 162
  spectroscopic monitoring 166
  total residuals 162
  variable residuals 162
MCR in practice 170
MCR-ALS 167
mean 248
  plot interpretation 189, 222
mean and Sdev
  plot interpretation 231
mean centering 248
mean normalization 73
Mean Square 148
mean-centering 80
median 248
median filtering 70, 71
minimize single responses 19
MixSum 248
Mixture Component 30, 31
mixture components 248
mixture constraint 248
mixture design 248
  PLS analysis 152
Mixture Design 248
  PLS analysis 152
mixture region 249

mixture response surface plot 155
mixture sum 249
Mixture Sum 249
mixture variable 249
mixture variables 17
MLR 107
  error measures 110
model 249
  check 151, 228
  constrained, non-mixture 153
  mixture 154
  robust 122
  validation 48
model center 249
model check 249
model distance 137
  plot interpretation 189
modeling power 137
  plot interpretation 189
modes 250
moving avaerage
  segment 259
moving average 70, 71
MS 148. *See* mean square
MSC 75
MSCorrection *See* MSC.
multi-linear constraint 250
Multi-Linear Constraint 250
multiple comparison tests 250
multiple comparisons 149
  plot interpretation 232
multiple linear regression 107. *See* MLR
multiplicative scatter correction 75
multivariate models
  validation 119
multivariate regression 105, 106
  model requirements 106
multi-way 250

N

noise 76, 250, 255
non-continuous variables 17. *See* category variables
non-design variables 17
  response variables 17
non-designed data 13
non-linearities 113, 151
non-linearity 251
non-negativity 251
normal distribution 251
  checking 251
normal probability plot 60, 251
normalization 72
  area 72
  maximum 73
  mean 73
  peak 73
  range 73
  unit vector 72
Norris-gap derivatives 76
n-plot 60
N-plot 60
n-plot of effects
  plot interpretation 226
n-plot of residuals
  plot interpretation 227
nPLS 262

O

O2V 52, 251
objective 16
offset 245, 251
one-way statistics 89
open file 55
optimal number of PCs 192, 195, 196
optimization 19, 251
orthogonal 251
orthogonal designs 252
outlier 99, 113, 252
  detect 217, 218, 219, 222, 227
  detect in PCA 99
  detect in regression 113
  influential 217, 218, 219
outlier detection 213
  prediction 233
outlier warnings 247
OV2 52, 252
overfitting 252

P

partial least squares 107. *See* PLS
passified 252
passify 82, 252
PCA 253
  interpret scores and loadings 99
  loadings 96
  purposes 93
  scores 96
  variances 95
PCA vs. curve resolution 94
PCR 13, 107, 253
PCs 94. *See* Principal Components
peak normalization 73
percentile 247, 248, 253, 264
percentiles 237
  interpretation 232
  plot interpretation 232
Plackett-Burman design 253
Plackett-Burman designs 22
planes 250

p-loadings 111
plot
  2D scatter 59
  2D scatter, raw data 62
  3D scatter 59
  3D scatter (raw data) 63
  contour 151
  histogram 61
  histogram (raw data) 64
  landscape 151
  line 58
  matrix 60
  matrix (raw data) 63
  normal probability 60
  normal probability (raw data) 64
  raw data, 2D scatter 62
  raw data, 3D scatter 63
  raw data, histogram 64
  raw data, line 61
  raw data, matrix 63
  raw data, normal probability 64
  response surface 151
  special plots 66
  stability 122
  table 67
  uncertainty 122
plot interpretation
  response surface, contour 224
  response surface, landscape 225
plot interpretation
  ANOVA 227
  bi-plot, scores and loadings 214
  box-plot 232
  classification scores 202
  classification table 228
  Cooman's plot 202
  cross-correlation (matrix plot) 225
  cross-correlation (table plot) 230
  detailed effects 185, 229
  discrimination power 185
  effects 226
  effects overview 229
  estimated concentrations 185
  estimated spectra 186
  f-ratios 186
  influence 203, 204, 211, 220
  interaction effects 230
  leverages 186, 222
  loading weights 189, 208, 209, 221
  loadings 187, 188, 205, 206, 207, 220, 221
  main effects 231
  mean 189, 222
  mean and Sdev 231
  model distance 189
  modeling power 189
  multiple comparisons 232
  percentiles 232
  predicted and measured 189
  predicted vs. measured 210, 230
  predicted vs. reference 211
  predicted with deviations 233
  prediction 230
  p-values of effects 190
  p-values of regression coefficients 190
  regression coefficients 190, 191, 223
  residuals 225, 227
  residuals vs. predicted 218
  residuals vs. scores 220
  response surface 224
  RMSE 192
  sample residuals 192, 193
  scatter effects 211
  scores 193, 212, 221
  Si vs. Hi 216
  Si/S0 vs. Hi 216
  standard deviation 194, 225
  standard errors 194
  total residuals 194
  variable residuals 197, 199, 201
  variance 195, 196, 197, 198, 199, 200, 201
  X-Y relation outliers 217
plots
  descriptive statistics 90
  normal probability 251
  various types 57
PLS 13, 107
  for constrained designs 152
  loading weights 111
  loadings 111
  scores 111
PLS discriminant analysis 138
PLS1 254
PLS2 254
precision 254
predicted and measured
  plot interpretation 189
predicted vs. measured
  plot interpretation 210, 230
predicted vs. Measured
  plot interpretation 210, 230
predicted vs. reference 132
  plot interpretation 211
predicted with deviation 132
predicted with deviations
  plot interpretation 233
predicted Y-values 110
prediction 131, 254
  allowed models 132
  in practice 133
  main results 132
  projection equation 131
  table plot interpretation 230

predictor 254
preference ratings
  plot as histogram 65
preprocessing 12
pre-processing 69
  three-way data 83
pre-treatment 69
primary objects 53
Primary Sample 254
Primary Variable 254
primary variables 53
principal component analysis 93
principal component regression 107. *See* PCR
principal components 94
principles of projection 94
print data 56
process variable 255
process variables 18
projection 94, 255
projection methods
  error measures 110
projection to latent structures 107. *See* PLS
proportional noise 255
pure components 256
p-value 148, 149, 150, 256
p-values of effects
  plot interpretation 190
p-values of regression coefficients
  plot interpretation 190

Q

q-loadings 111
quadratic effects 19
quadratic model 256
quadratic models 19

R

random effect 256
random order 256
random selection of test set 119, 120
randomization 43, 256
range normalization 73
ranges of variation
  how to select 47
raw data 12
  2D scatter plot 62
  3D scatter plot 63
  histogram 64
  line plot 61
  matrix plot 63
  n-plot 64
reference and center samples 149
reference sample 256
reference samples 42, 149
reflectance to absorbance 74
reflectance to Kubelka-Munk. 74
re-formatting 69
  fill missing 70
regression 105, 254, 257, 258
  multivariate 105, 106
  non-linearities 113
  outlier detection 113
  univariate 105, 106
regression coefficient 256
regression coefficients 109
  plot interpretation 190, 191, 223
  plot interpretation (tri-PLS) 191, 223
  uncertainty 122
regression methods 106, 112
regression modeling 114
  calibration 108
  validation 108
regression models
  shape 153
repeated measurement 257
replicate 257
replicates 42
residual 257
residual variance 95, 98, 257
residual variation 97
residual Y-variance 110
residuals 110, 245
  MCR 162
  n-plot 227
  plot interpretation 225
  sample 97
  variable 97
residuals vs. predicted
  plot interpretation 218
residuals vs. Scores
  plot interpretation 220
resolution 20, 22, 257
  fractional design 20, 22
response surface 246, 249
  mixture 155
  modeling 19
  plot interpretation 224
  plots 151
  results 150
response surface analysis 258
response surface modeling 150
response variable 258
response variables 16, 17
results
  clustering 145
  plot as histogram 66
  SIMCA 136
RMSE
  plot interpretation 192
RMSEC 110, 258

RMSED 258
RMSEP 110, 120, 258
root mean square error of prediction 120. *See* RMSEP
rotatability 23, 24

S

saddle point 151
sample 258
  residuals 97
sample distribution
  interpretation 213
sample leverage 137. *See* Hi
sample locations, interpretation 212
sample residuals
  MCR 162
  plot interpretation 192, 193
samples
  primary 53
  secondary 53
sample-to-model distance 137. *See* Si
Savitzky-Golay differentiation 76
Savitzky-Golay smoothing 70, 71
scaling 81, 259, 265
scatter effects 259
  plot interpretation 211
scores 96, 259
  plot interpretation 193, 212, 221
  PLS 111
  t 263
  t-scores 111
  u 264
  u-scores 111
scores and loadings
  bi-plot interpretation 214
screening 18, 259
  interaction effects 18
  interactions 18
  linear model 18
  main effects 18
screening designs 19
SDev 261
secondary objects 53
Secondary Sample 259
Secondary Variable 259
secondary variables 53
segment 259
segmented cross validation 120, 121
select
  ranges of variation 47
  regression method 112
  sesign variables 47
sensitivity to pure components 260
shift
  variables 80
Si 137
Si vs. Hi 138
  plot interpretation 216
Si/S0 vs. Hi
  plot interpretation 216
significance 121
significance level 260
significance testing 149
  center samples 149
  constrained designs 153
  COSCIND 149
  HOIE 149
  methods 149
  reference and center samples 149
  reference samples 149
significance testing methods 229
significance tests 112, 113
significant 260
significant effects
  detect 228, 229
SIMCA 135, 238, 260
  modeling 136
SIMCA classification 260
SIMCA results 136
  model results 136
  sample results 136, 137
  variable results 136
simplex 260
Simplex 28, 260
simplex-centroid design 260
simplex-lattice design 260
Singular Value Decomposition 107
smoothing 70
SNV 79
special plots 66
spectroscopic transformations 74
  absorbance to reflectance 74
  reflectance to absorbance 74
  reflectance to Kubelka-Munk 74
spectroscopy
  data 82
square effect 261
square root 70
SS 148
stability 122
stability plot
  segment information 124
standard designs 16
standard deviation 261
  plot interpretation 194, 225
standard errors
  plot interpretation 194
standard normal variate 79
standardization 81, 265
standardization of variables 261
star points distance to center 261
star samples 23, 261

distance to center 261
statistics
descriptive 89
descriptive plots 90
descriptive variable analysis 90
one-way 89
two-way 89
steepest ascent 262
student t-distribution 262
sum of squares 148
summary ANOVA 228

T

table plot 67
t-distribution 262
test samples 262
test set selection 119
group 119, 120
manual 119, 120
random 119, 120
test set validation 119, 262
tests of significance 112, 113
test-set switch 120
three-way 263
three-way data 51, 175, 235
counter-examples 179
examples 178
logical organization 52
modes 176
notation 176
OV2 and O2V 52
plot as matrix 64
pre-processing 83
ways 176
three-way PLS 13
three-way PLS Regression 262
three-way regression 179
total explained variance 98
total residual variance 98
total residuals
MCR 162
plot interpretation 194
training samples 262
transformations 69
averaging 80
derivatives 76
detect need 65
functions 70
logarithmic 70
MSC / EMSC 75
noise 76
shift variables 80
spectroscopic 74
standard normal variate SNV) 79
transposition 80
transpose 80
tri-PLS 13, 262
A-component model 180
inner relation 181
interpretation 182
loadings 180
main results 181
max number of PCs 182
one-component model 179
orthogonality 182
scores 180
weights 180, 181
X-variables 181
tri-PLS regression modeling 182
t-scores 111, 263
Tukey´s test 263
t-value 263
two-way statistics 89
types of experimental design 18

U

UDA 263
UDT 263
uncertainty limits 263
uncertainty test 121, 263
details 127
underfit 263
unfold 263
unfolding 3-D data 52
unimodality 263
unit vector normalization 72
univariate regression 105, 106
upper quartile 264
u-scores 111, 264
user-defined transformation 80

V

validation 94, 108, 241, 246, 262, 264
multivariate models 119
results 120
validation methods 119
cross validation 120
leverage correction 120
test set validation 119
Validation Methods 119
cross validation 120
leverage correction 120
test set validation 119
validation samples 264
variable 264
active 252
passified 252
residuals 97
variable residuals

MCR 162
plot interpretation 197, 199, 201
variables
primary 53
secondary 53
variance 265
degrees of freedom 242
explained 98
explained 95
interpretation 200
plot interpretation 195, 196, 197, 198, 199, 200, 201
residual 95, 98
stabilization 70
total explained 98
total residual 98
variances 95
variation 93
vertex sample 265

W

ways 265
weighting 81, 265
1/SDev 261
in PLS2 and PLS1 82
in sensory analysis 82
spectroscopy data 82
three-way data 83
weights
passify 252

X

X-Y relation outliers
plot interpretation 217
X-Y relationship
interpretation 207, 209
shape 218